Stata Implementation of the Non-Parametric Spatial Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator

P. Wilner Jeanty

The Kinder Institute for Urban Research
and
Hobby Center for the Study of Texas

Rice University

STATA CONFERENCE SAN DIEGO

July 26-27, 2012

Outline

Introduction

Background

- Researchers using geo-referenced data often need to contend with three critical issues:
  - Spatial correlation
  - Heteroskedasticity
  - Endogeneity
- These issues have been addressed from an econometric theory viewpoint (see for example, Conley, 1999; Kelejian and Prucha, 2007, 2010; Arraiz et al., 2010).
- However, they have often been overlooked in empirical applications.
- One reason is that estimators dealing with these conundrums are not always accessible.
- The purpose of this talk is to introduce two new user-written commands to implement the non-parametric spatial heteroskedasticity and autocorrelation consistent (SHAC) estimator of the variance covariance matrix in a spatial context.
- The SHAC estimator is robust against potential misspecification of the disturbance terms and allows for unknown forms of heteroskedasticity and correlation across spatial units.
- Heteroskedasticity is likely to arise when spatial units differ in size or in other structural features.

The Model

Model considered

$$y = X\beta + \gamma Y + \varepsilon \tag{1}$$

or more compactly

$$y = Z\delta + \varepsilon \tag{2}$$

with $Z = [X, Y]$ and $\delta = [\beta', \gamma']'$. Let H be an $n \times k_h$ matrix of instruments. The spatial covariance estimator in Conley (1999) is an application of Hansens (1982) generalized method of moments estimator (GMM) to spatial error autocorrelation. This estimator involves minimizing a quadratic form in the sample moment conditions, where the covariance matrix is obtained in non-parametric form a la Newey and West (1984). Specifically, the spatial covariances are estimated from weighted averages of sample covariances for pairs of observations that are within a given distance band from each other. Note that this approach requires covariance stationarity, which is only satisfied for a restricted set of spatial processes (e.g., it does not apply to spatial autoregressive (SAR) error models).

The GM Estimator

GM estimator

Based on the k_h-dimensional vector of instruments H, consider the following unconditional moment restrictions:

$$E_N[\psi(G_i, \delta)] = 0 \tag{3}$$

where E_N is the unconditional expectation operator over individuals and $\psi(G_i, \delta) = H_i^{'}(y_i - Z_i\delta)$. Corresponding to (3), the GMM estimator $\hat{\delta}$ for δ is the argument that minimizes

$$Q_N(\delta) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \psi(G_i, \delta) \right\}^{'} \Psi_N \left\{ \frac{1}{N} \sum_{i=1}^{N} \psi(G_i, \delta) \right\} \tag{4}$$

where Ψ_N is a positive definite matrix. The solution for the minimization problem in (4) is given by:

$$\hat{\delta}_{GMM} = \left(Z'H\Psi_N H'Z\right)^{-1} \left(Z'H\Psi_N H'y\right) \tag{5}$$

Let $\Psi_N = \hat{\Omega}^{-1}$. Provided that a consistent estimate $\hat{\Omega}$ of Ω can be obtained, the GMM estimator is efficient. In the spatial context, Conley (1999) suggests a procedure consistent with the Barlett window estimator proposed by Newey and West (1984).

Conley's SHAC Estimator

SHAC estimator

In particular, a consistent estimate $\hat{\Omega}$ of Ω to obtain standard errors robust to spatial autocorrelation and heteroskedasticity is given by:

$$\hat{\Omega} = N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} K(d_{ij}) \psi \left(G_i, \tilde{\delta} \right) \psi \left(G_i, \tilde{\delta} \right)' \qquad (6)$$

where $\tilde{\delta}$ is an estimate obtained in a first stage estimation such as two stage least squares and $K(d_{ij})$ is a weighting matrix. To ensure that $\hat{\Omega}$ is consistent and positive definite, the weighting matrix $K(d_{ij})$ is defined as the product of Barlett Kernels in two dimensions (North/South, East/West):

$$K(d_{ij}) = \left\{ \begin{array}{ll} (1 - d_{ij}^H / C_H)(1 - d_{ij}^V / C_V) & \text{if } d_{ij}^H < C_H \text{ and } d_{ij}^V < C_V \\ 0 & \text{otherwise} \end{array} \right\} \qquad (7)$$

where d_{ij}^H and d_{ij}^V represent the horizontal and vertical distances, respectively, between areal units i and j, and C_H and C_V represent the horizontal and vertical distance cutoffs beyond which no spatial correlation is assumed. The weights decline linearly from 1 to 0, ensuring the positive definiteness of $\hat{\Omega}$. Zero weights, thereby zero spatial autocovariances, result when one of the coordinates exceeds the distance cutoff. For more details, see Conley (1999). Once $\hat{\Omega}$ is obtained, the asymptotic variance-covariance of the parameter estimates can be derived.

Spatial Econometric Model

KP's model

The framework considered by Kelejian and Prucha (2007, hereafter KP) aims to accommodate spatial processes generated by Cliff-Ord type models. Inherent in these models are local nonstationarity and heteroskedasticity. Consider the following model:

$$y = X\beta + \lambda Wy + \gamma Y + \varepsilon \tag{8}$$

Equation (8) can be written in a compact form as

$$y = Z\delta + \varepsilon \tag{9}$$

with $Z = [X, Wy, Y]$ and $\delta = [\beta', \lambda, \gamma']'$.
In Kelejian and Prucha (2007) approach, the disturbance terms are assumed to follow a general spatial process of the form:

$$\varepsilon = R\xi \tag{10}$$

where ξ is a vector of i.i.d. (0, 1) innovations and R is an $n \times n$ non stochastic matrix whose elements are unknown and whose rows and column sums are uniformly bounded in absolute value.

KP's SHAC Estimator

SHAC estimation

As in Conley's case, the instrumental variable (IV) estimator of the parameters in equation (9) relies on a set of moment conditions of the form

$$EH'\varepsilon = 0 \tag{11}$$

The asymptotic distribution of the IV estimator will require the variance covariance matrix of the moment conditions defined by:

$$\Psi = VC(n^{-1/2}H'\varepsilon) = n^{-1}H'\Sigma H \tag{12}$$

where $\Sigma = RR'$ denotes the unknown variance covariance matrix of ξ. Let $\hat{\varepsilon} = y - Z\hat{\delta}_{S2SLS}$ and $\hat{\Psi}$ an estimate of Ψ. Kelejian and Prucha (2007) show that the (r,s) elements of $\hat{\Psi}$ can be consistently estimated by:

$$\hat{\Psi}_{r,s} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ir}h_{js}\hat{\varepsilon}_i\hat{\varepsilon}_j K(d^*_{ij}/d) \tag{13}$$

where the subscripts refer to the elements of the matrix of instruments H, d^*_{ij} is the distance between areal units i and j, $K()$ is a kernel function with the usual properties, d is the bandwidth or critical distance such that $K(d^*_{ij}/d) = 0$ for $d^*_{ij} \geq d$, and $\hat{\varepsilon}$ a vector of estimated residuals.

Asymptotic Distribution of $\hat{\delta}_{S2SLS}$

VC of parameter estimates

The choice of the bandwidth is more important than that of the kernel function (Cameron and Trivedi, 2005). In fact, so long as $K()$ is bounded, symmetric, real, and continuous, the kernel choice is immaterial (Mittelhammer et al., 2000). The bandwidth and the Kernel function place limits on the number of sample covariances. The bandwidth can be assumed either fixed or variable.
With $\hat{\Psi}$ available, the asymptotic variance covariance matrix of the spatial two-stage least squares estimates is given by:

$$\hat{\Phi} = n^2(\hat{Z}'\hat{Z})^{-1}Z'H(H'H)^{-1}\hat{\Psi}(H'H)^{-1}H'Z(\hat{Z}'\hat{Z})^{-1} \tag{14}$$

As a result, small sample inference concerning $\hat{\delta}_{S2SLS}$ can be based on the approximation $\hat{\delta}_{S2SLS} \sim N(\delta, n^{-1}\hat{\Phi})$.

Implementation Overview

Commands developed

- To implement the aforementioned SHAC estimators, we developed two Mata-based commands, `spcgmm` and `sphac`.
- `spcgmm` is essentially an estimation command.
- Since based on estimated residuals, `sphac` is a post-estimation command, though behaves as an estimation. command.
- Kelejian and Prucha (2007) allow the researcher to specify multiple distance measures. However, this version of `sphac` implements the SHAC estimator only in the case of a single distance measure.
- Both fixed and variable bandwidths are allowed.

Syntax for spcgmm

Command syntax

spcgmm *varlist* [if] [in], coord(*coordlist*) cutoff(*numlist*) [exog(*varlist*) endog(*varlist*) km level(#) collinear noconstant first]

Remarks

- When options `exog()` and `endog()` are not specified, the estimator becomes OLS with SHAC. OLS is a just-identified GMM estimator.
- Only the Barlett kernel is implemented

Syntax for sphac

Command syntax

sphac, dmat(*dmatrixname*) dfrom(*Mata*|*Stata*) [kernel(*functionname*) fbandw(#) vbandw(*varname*) noconst level(#) model(*ols*|*iv*|*sar*|*iv* − *sar*)]

Syntax for sphac

Command syntax

sphac, dmat(*dmatrixname*) dfrom(*Mata*|*Stata*) [kernel(*functionname*) fbandw(#) vbandw(*varname*) noconst level(#) model(*ols*|*iv*|*sar*|*iv* − *sar*)]

Kernel functions implemented

- Barlett: $K(z) = 1 - z$
- Epanechnikov: $K(z) = 1 - z^2$,
- Triangular: $K(z) = 1 - z$,
- Bisquare: $K(z) = (1 - z^2)^2$,
- Parzen: $K(z) = 1 - 6z^2 + 6|z|^3$ if $z \leq 0.5$ and $K(z) = 2(1 - |z|)^3$ if $0.5 < z \leq 1$.

Syntax for sphac

Command syntax

sphac, dmat(*dmatrixname*) dfrom(*Mata*|*Stata*) [kernel(*functionname*) fbandw(#) vbandw(*varname*) noconst level(#) model(*ols*|*iv*|*sar*|*iv* − *sar*)]

Kernel functions implemented

- Barlett: $K(z) = 1 - z$
- Epanechnikov: $K(z) = 1 - z^2$,
- Triangular: $K(z) = 1 - z$,
- Bisquare: $K(z) = (1 - z^2)^2$,
- Parzen: $K(z) = 1 - 6z^2 + 6|z|^3$ if $z \leq 0.5$ and $K(z) = 2(1 - |z|)^3$ if $0.5 < z \leq 1$.

Requirements

- sphac requires a pre-calculated distance matrix and a pre-generated variable holding distance to nearest neighbors when users specify the `vband()` option. This can be done easily using the user-written command nearstat.
- sphac also uses saved results from estimation commands to perform all calculations. So far, it works after the official Stata commands `regress` and `ivregress` and after the user-written command `spivreg`.

Data

Data description

- Examples use a dataset of 1789 Census tracts for the State of Michigan.
- Variables include:
- Dependent
  - Change in log population 1990 – 2000 (popch)
- Independent
  - Racial diversity, 2000 (divx) - Assumed to be endogenous
  - Log population, 1990 (lnpop9)
  - College graduate, 1990 (bspct9)
  - Median household income, 1990 (lavhhin9)
  - Unemployment rate, 1990 (unemprt9)
  - Employment share in agriculture, 1990 (pctfarm9)

```
. use michigan_tracts, clear
. global xvars lnpop9 bspct9 lavhhin9 unemprt9 pctfarm9
```

Data Summary

Data description

```
. summarize popch divx $xvars, separator(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
popch	1789	.051171	.2530615	-2.241498	2.489235
divx	1789	.2667414	.184348	.0283146	.8802574
lnpop9	1789	8.071044	.4616808	4.927254	9.167328
bspct9	1789	11.97815	10.29886	0	62.67878
lavhhin9	1789	10.54695	.4205478	8.966855	12.31559
unemprt9	1789	9.249566	8.310277	0	52.37288
pctfarm9	1789	.9547646	1.264445	0	12.14511

Racial Diversity: The variable of interest

Aspects of racial diversity

- Racial diversity is assumed to be endogenous due to reverse causation, as migration affects the spatial distribution of the minority populations. Also, political leaders may pursue policies that influence diversity.
- There are pros and cons of racial diversity.
- Opponents vehemently maintain that racial diversity may cause conflict of preferences, racism, and prejudices that are often conducive to counter-productive policies for society as a whole.
- Proponents forcefully argue that ethnic diversity propels variety in skills, experiences that lead to innovations and creativity.
- Communities clinging to these views may implement anti or pro-diversity policies that ward off or attract migrants.
- The variable racial diversity, defined as the Theil's entropy index, was calculated using block group level data for four ethnic groups: Hispanic, Non-Hispanic White, Non-Hispanic Black, and Non-Hispanic Asian.

$$divx = \sum_{m=1}^{M} \pi_m ln(1/\pi_m) \tag{15}$$

where m indexes the ethnic groups and π_m is the share of the ethnic group m in a census tract.

Spatial Interactions and Spatial Weights

Rationale for spatial interactions

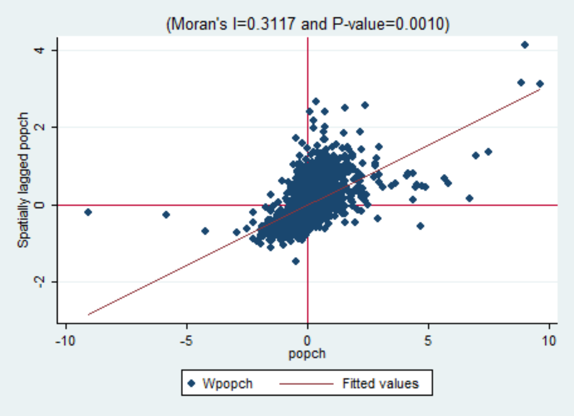
- Growing or declining neighborhoods tend to be located near each other in geographic space because they generally have similar access to transportation, zoning, and topography that supports housing construction.
- Also, economic shocks affecting migration decisions may be transmitted across borders, or a community is attracting migrants simply because its neighbors are doing so.
- As a result, some spillover effects across geographically proximate neighborhoods are expected.
- To get a sense of the spatial distribution of population growth, I constructed a Moran scatterplot by coding:

```
. splagvar popch, wname(winvsq) wfrom(Mata) moran(popch) plot(popch)
(permute popch : splagvar_randper)
```
(*output omitted*)

- The spatial weights matrix `winvsq` was generated using the user-written command `spwmatrix`.

Moran Scatterplot for Population Growth

Plot from splagvar



Model Estimation

Instrumental variables

- Given that both diversity and population growth use population data, it is difficult to find instruments that are correlated with diversity but uncorrelated with shocks to population growth.
- In this exercise, estimations will rely on three constructed instruments.
- A quasi-instrument, `q_divx`, was generated using the user-written command `splagvar` as follows

```
. qui splagvar, qvar(divx) qname(q_divx)
```

- This variable takes on the values of -1, 0, and 1 if the values of `divx` are in the bottom, middle, and top third respectively (Fingleton and Le Gallo, 2008).
- Two other instruments was constructed by data transformation based on the notion that if the endogenous regressor Y has a skewed distribution, the following transformation of the data may yield valid instruments Lewbel (1997):

$$\begin{aligned} liv1 &= (y_i - \bar{y})(Y_i - \bar{Y}) \\ liv2 &= (Y_i - \bar{Y})^2 \end{aligned} \tag{16}$$

Demonstration of spcgmm

Estimation procedures

- To implement the Conley's procedure, a distance cutoff is needed. Researchers usually use 10 miles when working with Census tract level data (.eg., (Jeanty et al., 2010; Boarnet et al., 2003). We use 8.58 miles implied by distances to first nearest neighbors calculated using the user-written command `nearstat`. This will be the first model estimated.

nearstat output

```
. nearstat (intptlat intptlon), near(intptlat intptlon) distv(neardist1) ///
> r(3958.761) des(stat)
```

Descriptive Statistics for Distance

Variable	Obs	Mean	Std	Min	Max
distance*	3198732	57.20	46.43	0.23	198.75
neardist1**	1789	1.21	1.16	0.23	8.57

```
*:   Distance between each input feature and all near features
**:  Distance from each input feature to its first nearest neighbor

 Distance (in miles) calculations completed successfully and/or all requests
> processed
```

GMM Estimation

spcgmm output

```
. spcgmm popch $xvars, exog(q_divx liv1 liv2) endog(divx)  ///
> coord(intptlat intptlon) cutoff(8.58 8.58)
Spatial 2-Step GMM (Mata version)
                        Number of observations = 1789
                        Crit. fnct. test of overid. restrictions =   1.4788842
                        DF= 2
                        P-value = 0.47738
------------------------------------------------------------------------------
       popch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        divx |   .1671914       .0403    4.15   0.000     .0881511    .2462317
      lnpop9 |  -.1673229    .0350197   -4.78   0.000     -.236007   -.0986389
      bspct9 |  -.0046652     .001249   -3.74   0.000    -.0071149   -.0022155
    lavhhin9 |   .1943346    .0394714    4.92   0.000     .1169195    .2717497
    unemprt9 |  -.0070459    .0015382   -4.58   0.000    -.0100627   -.0040291
    pctfarm9 |   .0319771    .0060263    5.31   0.000     .0201577    .0437964
       _cons |  -.6007922    .4340829   -1.38   0.167    -1.452157    .2505729
------------------------------------------------------------------------------
Instrumented:   divx
Instruments:    lnpop9 bspct9
                lavhhin9 unemprt9
                pctfarm9 q_divx liv1
                liv2

. eststo
(est2 stored)
```

Demonstration of sphac

Estimation procedures

- The demonstration of `sphac` makes use of the outstanding user-written `spivreg` command (Drukker et al., 2011), which requires spatial weights in Mata memory.
- A forthcoming updated version of the user-written command `spwmatrix` has an external option allowing one to store spatial weights as a Mata object residing in Mata memory. For this demonstration, we use two spatial weights, `winvsq` and `wcontig`.
- `winvsq`, an inverse distance squared matrix, was generated by `spwmatrix`, but `wcontig`, a contiguity matrix, was created in `ArcGIS` and imported into Stata by `spwmatrix`. Both spatial weights matrices were then stored as Mata objects for the estimations.

Demonstration of sphac

Estimation procedures

- Based on Kelejian and Prucha (2007, Assumption 4a), the number of neighbors within the bandwidth is constrained by $l_n = o(n^{1/3})$.
- This yields a threshold number of 12 neighbors. We will use a variable bandwidth corresponding to distance to the 12th nearest neighbor for each observation.
- Next steps consist in calculating distance to the 12^{th} nearest neighbors and in storing the distance matrix to a Mata file.
- Three more models will then be estimated.
- Model 2 allows for spatial interactions and is estimated by spatial two-stage least squares.
- Model 3 is also estimated by spatial 2SLS but with Parzen kernel SHAC standard errors. The Barlett kernel yields similar results up to 3 decimal digits.
- As an alternative to model 3, model 4 allows for heteroskedastic innovations and disturbances that follow a first order autoregressive process:

$$\varepsilon = \rho W + \xi \tag{17}$$

- Kelejian and Prucha (2010) argue that model 3 is more robust than model 4.

Distance to 12 nearest neighbors

nearstat output

```
. nearstat (intptlat intptlon), near(intptlat intptlon) distv(neardist12) ///
> kth(12) r(3958.761) des(stat) expdist(distmat) expto(Mata)
```

Descriptive Statistics for Distance

Variable	Obs	Mean	Std	Min	Max
distance*	3198732	57.20	46.43	0.23	198.75
neardist12**	1789	3.57	3.03	1.03	24.42

```
*:   Distance between each input feature and all near features
**:  Distance from each input feature to its 12th nearest neighbor

 Distance (in miles) calculations completed successfully and/or all requests
> processed

 Also, distance between input and near features exported to the Mata file:
> C:\data\Stata_Conference2012/distmat.

. gen neardist12a=neardist12+0.01 // To guarantee 12 neighbors for each
>observation
```

Spatial Two-Stage Least Squares

spivreg output

```
. spivreg popch (divx=q_divx liv1 liv2) $xvars, id(obsid_n) dlmat(winvsq)
Spatial autoregressive model                    Number of obs    =     1789
(GS2SLS estimates)
```

popch	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
popch						
divx	.1441401	.0295278	4.88	0.000	.0862667	.2020135
lnpop9	-.1140855	.0107016	-10.66	0.000	-.1350603	-.0931108
bspct9	-.0027862	.0007404	-3.76	0.000	-.0042373	-.0013351
lavhhin9	.101096	.0254452	3.97	0.000	.0512244	.1509676
unemprt9	-.0046386	.0009612	-4.83	0.000	-.0065226	-.0027546
pctfarm9	.0115774	.0042329	2.74	0.006	.0032812	.0198737
_cons	-.0905023	.2806597	-0.32	0.747	-.6405853	.4595807
lambda						
_cons	.6388438	.065243	9.79	0.000	.5109698	.7667178

```
Instrumented:  divx
Instruments:   q_divx liv1 liv2

. eststo
(est2 stored)
```

Spatial Two-Stage Least Squares with SHAC

sphac output

```
. sphac, dmat(distmat) dfrom(Mata) vbandw(neardist12a) kernel(Parzen) ///
> model(iv-sar)

Spatial HAC Standard Errors
Kernel = Parzen
Bandwidth =  Variable
```

popch	Coef.	SHAC Std. Err.	z	P>|z|	[95% Conf.	Interval]
popch						
divx	.1441401	.029667	4.86	0.000	.085994	.2022863
lnpop9	-.1140855	.0323895	-3.52	0.000	-.1775678	-.0506032
bspct9	-.0027862	.000841	-3.31	0.001	-.0044344	-.0011379
lavhhin9	.101096	.0314031	3.22	0.001	.0395471	.1626449
unemprt9	-.0046386	.0011428	-4.06	0.000	-.0068783	-.0023988
pctfarm9	.0115774	.0043942	2.63	0.008	.002965	.0201899
_cons	-.0905023	.3568791	-0.25	0.800	-.7899723	.6089678
lambda						
_cons	.6388438	.0925785	6.90	0.000	.4573932	.8202944

```
Instrumented:  divx
Instruments:   lnpop9 bspct9
               lavhhin9 unemprt9
               pctfarm9 q_divx liv1
               liv2

. eststo
(est3 stored)
```

Generalized Spatial Two-Stage Least Squares

spivreg output

```
. spivreg popch (divx=q_divx liv1 liv2) $xvars, id(obsid_n) dlmat(winvsq) ///
> elmat(wcontig) het nolog

Spatial autoregressive model                    Number of obs   =      1789
(GS2SLS estimates)
```

popch	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
popch						
divx	.1624854	.0323197	5.03	0.000	.09914	.2258309
lnpop9	-.1065138	.0298879	-3.56	0.000	-.1650931	-.0479346
bspct9	-.0027052	.0009132	-2.96	0.003	-.004495	-.0009154
lavhhin9	.0908998	.0326279	2.79	0.005	.0269502	.1548493
unemprt9	-.0043872	.001261	-3.48	0.001	-.0068588	-.0019156
pctfarm9	.006457	.0044147	1.46	0.144	-.0021956	.0151097
_cons	-.0518782	.3650912	-0.14	0.887	-.7674437	.6636874
lambda						
_cons	.7343	.0912013	8.05	0.000	.5555487	.9130512
rho						
_cons	.1316497	.0815937	1.61	0.107	-.028271	.2915705

```
Instrumented:  divx
Instruments:   q_divx liv1 liv2

. eststo
(est4 stored)
```

Comparison of Results

Regression outputs

Table : Regression Results across Estimation Methods

	GMM W/ SHAC	S2SLS	S2SLS W/ SHAC	GS2SLS HET
Racial div. 2000	0.1672***	0.1441***	0.1441***	0.1625***
	(0.0403)	(0.0295)	(0.0297)	(0.0323)
Log pop. 1990	-0.1673***	-0.1141***	-0.1141***	-0.1065***
	(0.0350)	(0.0107)	(0.0324)	(0.0299)
Col. grad. 1990	-0.0047***	-0.0028***	-0.0028***	-0.0027***
	(0.0012)	(0.0007)	(0.0008)	(0.0009)
Log inc. 1990	0.1943***	0.1011***	0.1011***	0.0909***
	(0.0395)	(0.0254)	(0.0314)	(0.0326)
Unempl. 1990	-0.0070***	-0.0046***	-0.0046***	-0.0044***
	(0.0015)	(0.0010)	(0.0011)	(0.0013)
Agr. jobs 1990	0.0320***	0.0116***	0.0116***	0.0065
	(0.0060)	(0.0042)	(0.0044)	(0.0044)
Intercept	-0.6008	-0.0905	-0.0905	-0.0519
	(0.4341)	(0.2807)	(0.3569)	(0.3651)
lambda		0.6388***	0.6388***	0.7343***
		(0.0652)	(0.0926)	(0.0912)
rho				0.1316
				(0.0816)
N	1789	1789	1789	1789

Standard errors in parentheses
* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Final thoughts

Summary and observation

- In this presentation, we illustrate two new user-written commands, spcgmm and sphac.
- We show how three typical econometric issues endogeneity, spatial autocorrelation, and heteroskedasticity facing researchers using geo-referenced data can be addressed in Stata.
- In the contrived examples, we estimated a population growth model with racial diversity as the explanatory variable of interest.
- The results show that, net of economic and demographic factors, racial diversity is positively correlated with population growth.

Final thoughts

Summary and observation

- In this presentation, we illustrate two new user-written commands, spcgmm and sphac.
- We show how three typical econometric issues endogeneity, spatial autocorrelation, and heteroskedasticity facing researchers using geo-referenced data can be addressed in Stata.
- In the contrived examples, we estimated a population growth model with racial diversity as the explanatory variable of interest.
- The results show that, net of economic and demographic factors, racial diversity is positively correlated with population growth.

Limitations and potential improvements

- Implementation of `sphac` depends on a dense, rather than sparse, distance matrix
- Large sample size may be a problem, though the command work well for US county level data.
- Improvements will depend on the availability of sparse matrix operations in Mata.

Final thoughts

Summary and observation

- In this presentation, we illustrate two new user-written commands, spcgmm and sphac.
- We show how three typical econometric issues endogeneity, spatial autocorrelation, and heteroskedasticity facing researchers using geo-referenced data can be addressed in Stata.
- In the contrived examples, we estimated a population growth model with racial diversity as the explanatory variable of interest.
- The results show that, net of economic and demographic factors, racial diversity is positively correlated with population growth.

Limitations and potential improvements

- Implementation of `sphac` depends on a dense, rather than sparse, distance matrix
- Large sample size may be a problem, though the command work well for US county level data.
- Improvements will depend on the availability of sparse matrix operations in Mata.

Next steps

- We will write the help files and submit to SSC
- Finally, we will consider extend sphac to make it work after non-linear models.

Thank you!!!

References

Arraiz, I., Drukker, D. M., Kelejian, H. H., and Prucha, I. R. (2010). A spatial cliff-ord-type model with heteroscedastic innovations: Small and large sample results. *Journal of Regional Science*, 50:592–614.

Boarnet, M., Chalermpong, S., and Geho, E. (2003). Specification issues in models of population and employment growth. *Papers in Regional Science*, 84:21–46.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, 92:1–45.

Drukker, D. M., Prucha, I. R., and Raciborski, R. (2011). A command for estimating spatial-autoregressive models with spatial autoregressive disturbances and additional endogenous variables. University of Maryland Working paper.

Fingleton, B. and Le Gallo, J. (2008). Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: Finite sample properties. *Papers in Regional Science*, 87(3):319–339.

Jeanty, P. W., Partridge, M. D., and Irwin, E. (2010). Estimation of a spatial simultaneous equation model of population migration and housing price dynamics. *Regional Science and Urban Economics*, pages 343–352.

Kelejian, H. H. and Prucha, I. R. (2007). Hac estimation in a spatial framework. *Journal of Econometrics*, 140:131–154.

Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroscedastic disturbances. *Journal of Econometrics*, 157:53–67.

Lewbel, A. (1997). Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and r&d. *Econometrica*, 65:1201–1213.

Mittelhammer, R., Judge, G., and Miller, D. (2000). *Econometric Foundations*. Cambridge University Press, Cambridge.

Newey, W. and West, K. (1984). A simple, positive semi-definite, heteroskedastic and autocorrelated consistent covariance matrix. *Econometrica*, 55:703–708.