An introdution to spatial econometrics using Stata

David M. Drukker

Executive Director of Econometrics
Stata

2018 Mexican Stata Users Group meeting
16–17 August 2018

What is spatial econometrics?

- Suppose we have cross-sectional data on individuals $i \in \{1, 2, \ldots, N\}$
  - In standard econometrics/stastics we assume that the outcomes of any two individuals y_i and y_j are idependent, after conditioning on covariates
  - In spatial econometrics/statistics, we allow the outcomes of any two individuals y_i and y_j to depend on each other, after conditioning on covariates
    - The dependence could be because the outcome of person i functionally affects the outcome of person j
    - The dependence could be because the errors that drive the outcome of person i are correlated with the errors that drive the outcome of person j

What is spatial econometrics?

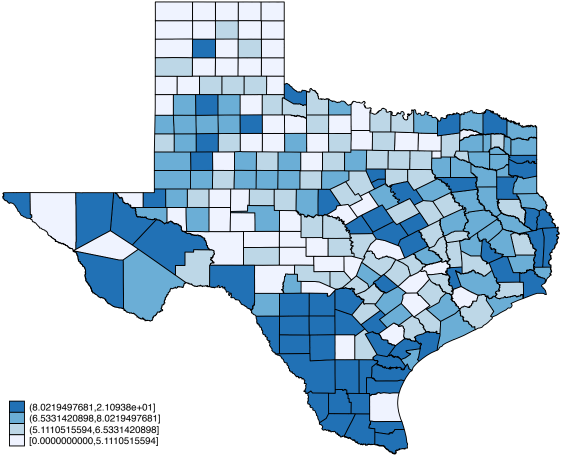
- Spatial econometrics/statistics is a class of estimation and inference methods for models in which the individual outcomes depend on each other
- Individuals could be people, places, firms, ...
- The data could be panel data or longitudinal data with many individuals and a fixed number of time periods
- The data do not have be geographic to be spatial
    - Network relationships can be modeled using this framework

Literature

- For an introduction to Spatial econometrics and many citations to the original literature see
  https://www.stata.com/manuals/sp.pdf
- The GS2SLS estimator was derived by Kelejian and Prucha (1998, 1999, and 2010)
  It was extended by Arraiz et al. (2010), Drukker et al. (2013a) Drukker, Prucha, and Raciborski (2013c and 2013d) and Drukker, Peng, Prucha, and Raciborski (2013b), provide an introduction to spatial econometrics and discuss implementation details for GS2SLS and maximum-likehood (ML) estimation
- Lee (2004) derives the ML estimator and the robust VCE of the QML estimator
- For panel data, see Lee and Yu (2010a, 2010b), and Kapoor, Kelejian, and Prucha (2007)

Geographic example: Unemployment rates in Texas

```
. use texas_unemp, clear
. grmap unemployment
```



Spatial autoregressive model for unemployment

$$unemp_i = \lambda \sum_{j=1}^{n} w_{i,j} unemp_j + \mathbf{x}_i \beta' + \epsilon_i$$

$$\mathbf{x}_i = (gini_i, divorce_i, age_i, lnpdensity_i, constant)$$

- Unemployment in place i depends on a weighted average of unemployment in the other places and a linear function of covariates
- The weights $w_{i,j}$ are given,
  - they are part of the model
  - they parameterize how important, or close to, each individual is to every other individual

Spatial autoregressive model for unemployment

- It helps to write

$$unemp_i = \lambda \sum_{j=1}^{n} w_{i,j} unemp_j + \mathbf{x}_i \beta' + \epsilon_i$$

$$\mathbf{x}_i = (gini_i, divorce_i, age_i, lnpdensity_i, constant)$$

in vector form

$$\mathbf{unemp} = \lambda \, \mathbf{W} \, \mathbf{unemp} + \mathbf{X}\beta' + \boldsymbol{\epsilon}$$

- **unemp** is $N \times 1$ vector of observations on *unemp*
- **W** is an $N \times N$ matrix of weights – spatial weighting matrix and $\mathbf{W}[i,j] = w_{i,j}$
- **X** is $N \times k$ vector of observations on the covariates
- $\boldsymbol{\epsilon}$ is $N \times k$ vector of errors

- View

$$\mathbf{unemp} = \lambda\ \mathbf{W}\ \mathbf{unemp} + \mathbf{X}\beta' + \epsilon$$

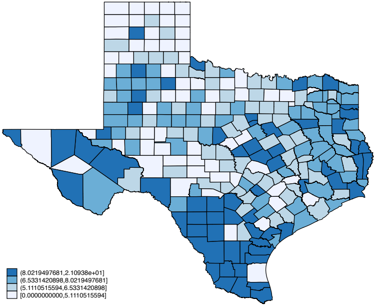
as an $N \times 1$ system of equations for **unemp**

- The term λ **W unemp** is known as a spatial lag of the dependent variable
- Things about the spatial weighting matrix **W**
    - The elements of **W** parameterize who is close whom
    - The diagonal elements are zero
      The unemployment level in place i does not affect itself
    - In a sense, the scale does not matter
      λ models the scale, multiplying **W** by a scalar does not matter
    - In a sense, the scale does matter
      If $\lambda\mathbf{W}$ is too large, the system is not stable
      Normalize **W** by its largest eigenvalue for a natural measure of when λ is too large

- The equation

$$\textbf{unemp} = \lambda\ \mathbf{W}\ \textbf{unemp} + \mathbf{X}\beta' + \epsilon$$

  says that dark regions are clustered together and light regions are cluster together, because the unemployment level in place *i* functionally affects the unemployment level in near by places
- "near by" is parameterized by **W**
- **W** is fixed, λ is estimated



Four important equations

- Solving

$$\textbf{unemp} = \lambda\ \textbf{W}\ \textbf{unemp} + \textbf{X}\boldsymbol{\beta}' + \boldsymbol{\epsilon} \tag{1}$$

for **unemp** yields

$$\textbf{unemp} = (\textbf{I} - \lambda\ \textbf{W})^{-1}\textbf{X}\boldsymbol{\beta}' + (\textbf{I} - \lambda\ \textbf{W})^{-1}\boldsymbol{\epsilon} \tag{2}$$

$\textbf{I}$ is $N \times N$ identity matrix

- From equation (2), the mean of **unemp** given covariates $\textbf{X}$ is

$$\textbf{E}[\textbf{unemp}|\textbf{X}] = (\textbf{I} - \lambda\ \textbf{W})^{-1}\textbf{X}\boldsymbol{\beta}' \tag{3}$$

- From equation (3), the conditional mean of the unemployment level in place i can be written as

$$\textbf{E}[unemp_i|\textbf{X}] = s_{i,i}\textbf{x}_i\boldsymbol{\beta} + \sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\textbf{x}_j\boldsymbol{\beta} \tag{4}$$

where $s_{i,j}$ is the (i,j) element of $(\textbf{I} - \lambda\ \textbf{W})^{-1}$

Direct and indirect effects

$$\mathbf{E}[unemp_i|\mathbf{X}] = s_{i,i}\mathbf{x}_i\boldsymbol{\beta} + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\mathbf{x}_j\boldsymbol{\beta} \tag{4}$$

- The first term in equation (4) gives rise to the direct effect of a covariate on the outcome
  - In the first term, the covarates of observation in i only affect the unemployment level in place i
  - So a change in the k(th) covariate from observation i has a direct effect on the outcome in place i
    This effect is also known as an "own" effect, because the change in a covariate in place i affect the outcome in the same place

Direct and indirect effects

$$\mathbf{E}[unemp_i|\mathbf{X}] = s_{i,i}\mathbf{x}_i\boldsymbol{\beta} + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\mathbf{x}_j\boldsymbol{\beta} \tag{4}$$

- The second term in equation (4) gives rise to the indirect effects of a covariate on the outcome
  These effects effects are also known as spill-over effects
  - In the second term, the covariates of observations in $j \neq i$ affect the unemployment level in place i
  - So changes in the k(th) covariate from observations $j \neq i$ have indirect effects on the outcome in place i
    These effects are also known as "spill-over" effects, because the change in a covariate in place $j \neq i$ "spills over" to affect the outcome in a different place

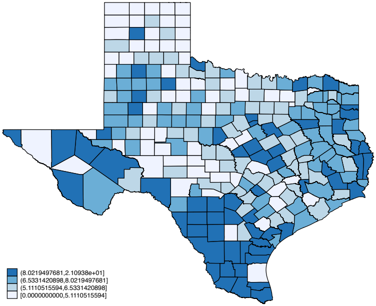
Direct and indirect effects

$$\mathbf{E}[\mathbf{unemp}|\mathbf{X}] = (\mathbf{I} - \lambda\ \mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}'$$

$$\mathbf{E}[unemp_i|\mathbf{X}] = s_{i,i}\mathbf{x}_i\boldsymbol{\beta} + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\mathbf{x}_j\boldsymbol{\beta}$$

- When $\lambda = 0$
  $(\mathbf{I} - \lambda\ \mathbf{W})^{-1} = \mathbf{I}$
- When $\lambda = 0$
  $s_{i,i} = 1$ and for $(j \neq i)$ $s_{i,j} = 0$

- Looking at



- I want $unemp_j$ to have a weight of 1 in the equation for $unemp_i$ if places j and i share a boundary and to have a weight of 0 otherwise
- In other words, I want $\mathbf{W}$ to be a normalized contiguity matrix
  - A contingguity matrix is a matrix of zeros and ones

$$\mathbf{W}[i,j] = \begin{cases} 1 & \text{if i shares a boundary with j} \\ 0 & \text{otherwise} \end{cases}$$

A model for unemployment

- I have my analysis data in `texas_unemp`.
- I have already used `spset` to link `texas_unemp` with the shapefile data in `texas_county`
- In the next section of the talk, I will go through the details of this process
  - First, I am going to analysis this data and show you what you learn from it
  - Later, I show the boring details about how to set up the data

```
. clear all
. use texas_unemp, clear
. spset
  Sp dataset texas_unemp.dta
                data:  cross sectional
     spatial-unit id:  _ID
         coordinates:  _CX, _CY (planar)
    linked shapefile:  tl_2016_us_county_shp.dta
```

A model for unemployment

- Now that I have my `spset` data in memory, I create a normalized contiguity matrix for the Texas counties named `C`

```
. spmatrix create contiguity C
```

- I use `spregress , gs2sls` to estimate the parameters of

$$\textbf{unemp} = \lambda \ \mathbf{W} \ \textbf{unemp} + \mathbf{X}\beta' + \epsilon$$

by generalized spatial two stage least squares (GS2SLS)

```
. spregress unemployment gini divorce age ln_pdensity , dvarlag(C) gs2sls
  (254 observations)
  (254 observations (places) used)
  (weighting matrix defines 254 places)
Spatial autoregressive model                    Number of obs     =        254
GS2SLS estimates                                Wald chi2(5)      =     252.59
                                                Prob > chi2       =     0.0000
                                                Pseudo R2         =     0.4966
```

unemployment	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
unemployment						
gini	.523163	.0452371	11.56	0.000	.4345	.611826
divorce	-.0872785	.1082299	-0.81	0.420	-.2994053	.1248483
age	-.137939	.0336929	-4.09	0.000	-.2039759	-.0719021
ln_pdensity	.6253197	.1015189	6.16	0.000	.4263463	.8242931
_cons	-10.91913	2.212596	-4.93	0.000	-15.25574	-6.582519
C						
unemployment	.022791	.0804118	0.28	0.777	-.1348132	.1803953

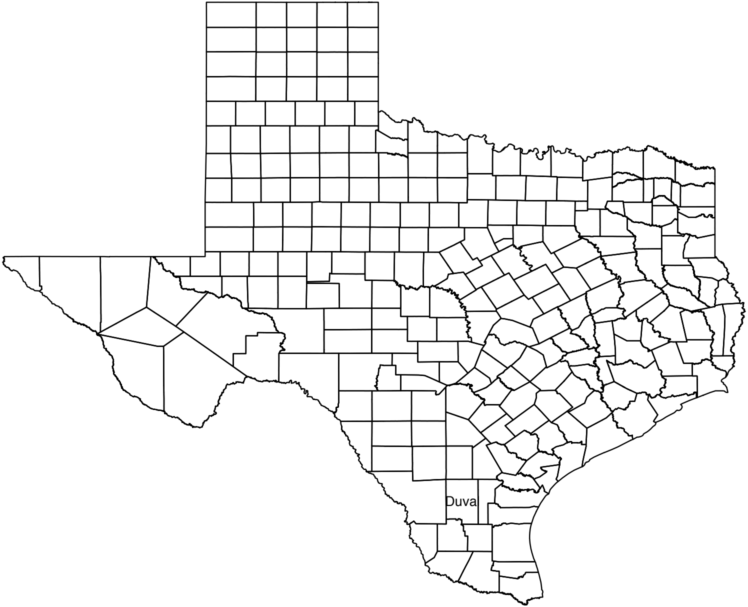
```
Wald test of spatial terms:          chi2(1) = 0.08        Prob > chi2 = 0.7768
```

- I use `spregress, ml` to estimate the parameters of

$$\mathbf{unemp} = \lambda\ \mathbf{W}\ \mathbf{unemp} + \mathbf{X}\beta' + \epsilon$$

by quasi maximum likelihood

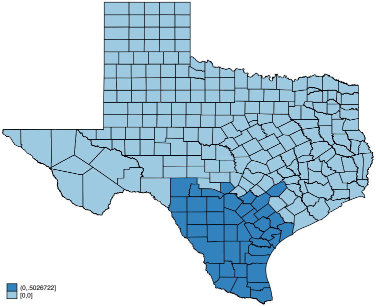
```
. spregress unemployment gini divorce age ln_pdensity , dvarlag(C) ml nolog
  (254 observations)
  (254 observations (places) used)
  (weighting matrix defines 254 places)
Spatial autoregressive model                    Number of obs     =        254
Maximum likelihood estimates                    Wald chi2(5)      =     260.27
                                                Prob > chi2       =     0.0000
Log likelihood = -544.67449                     Pseudo R2         =     0.4888

------------------------------------------------------------------------------
unemployment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
unemployment |
        gini |   .5011175   .0440013    11.39   0.000     .4148766    .5873585
     divorce |  -.0763862   .1073923    -0.71   0.477    -.2868713    .1340988
         age |  -.1321324   .0333759    -3.96   0.000    -.1975479    -.066717
 ln_pdensity |    .613427   .1007021     6.09   0.000     .4160544    .8107995
       _cons |  -10.91903    2.19738    -4.97   0.000    -15.22581    -6.61224
-------------+----------------------------------------------------------------
C            |
unemployment |    .134576   .0652929     2.06   0.039     .0066044    .2625477
-------------+----------------------------------------------------------------
var(e.unem~t)|   4.256417   .3778386                       3.57671    5.065294
------------------------------------------------------------------------------
est of spatial terms:            chi2(1) = 4.25         Prob > chi2 = 0.0393
```


Duva

A change to one place affects near-by places

$$\mathbf{E}[\mathbf{unemp}|\mathbf{X}] = (\mathbf{I} - \lambda\ \mathbf{W})^{-1}\mathbf{X}\beta'$$

```
. predict yhat0 , rform
. generate double gini_orig = gini
. replace gini = gini + 1 if cname=="Duval"
(1 real change made)
. predict yhat1 , rform
. generate diff = yhat1 - yhat0
. replace  gini = gini_orig
(1 real change made)
. grmap diff
```



$$\mathbf{E}[unemp_i|\mathbf{X}] = s_{i,i}\mathbf{x}_i\boldsymbol{\beta} + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\mathbf{x}_j\boldsymbol{\beta}$$

$$\frac{\partial \mathbf{E}[unemp_i|\mathbf{X}]}{\partial \mathbf{x}_k} = s_{i,i}\beta_k + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k$$

- Note that we are changing the value of covariate k in all places ($\partial \mathbf{x}_k$ instead of ∂x_k)
- Note that β_k is neither the direct nor the indirect effect
- marginal effect on the mean outcome in observation i of an infinitesimal change in each observation on covariate k
- This effect is for the place i
  - There N effects
  - Estimate the mean of these N effects

$$\frac{\partial \mathbf{E}[unemp_i|\mathbf{X}]}{\partial \mathbf{x}_k} = s_{i,i}\beta_k + \sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\beta_k$$

$$\text{Average of total effects} = 1/n \sum_{i=1}^{n} \left(s_{i,i}\beta_k + \sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\beta_k \right)$$

$$\text{Average of direct effects} = 1/n \sum_{i=1}^{n} (s_{i,i}\beta_k)$$

$$\text{Average of indirect effects} = 1/n \sum_{i=1}^{n} \left(\sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\beta_k \right)$$

- Use `estat impact` to estimate the means of the direct efffect, the indirect effects, and the total effects of a marginal (derivative) change in each covariate k

```
. estat impact
progress   : 25%  50%  75% 100%
Average impacts                              Number of obs     =        254

------------------------------------------------------------------------------
             |              Delta-Method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
direct       |
        gini |   .5023771   .0437608    11.48   0.000     .4166074    .5881467
     divorce |  -.0765782   .1076511    -0.71   0.477    -.2875704     .134414
         age |  -.1324646   .0334277    -3.96   0.000    -.1979817   -.0669474
  ln_pdensity|   .6149688   .1008609     6.10   0.000     .4172852    .8126524
-------------+----------------------------------------------------------------
indirect     |
        gini |   .0655407   .0344884     1.90   0.057    -.0020554    .1331368
     divorce |  -.0099905   .0147609    -0.68   0.499    -.0389213    .0189403
         age |  -.0172815   .0099729    -1.73   0.083    -.0368279     .002265
  ln_pdensity|   .0802296   .0448002     1.79   0.073    -.0075772    .1680363
-------------+----------------------------------------------------------------
total        |
        gini |   .5679178   .0526802    10.78   0.000     .4646664    .6711692
     divorce |  -.0865687   .1215044    -0.71   0.476    -.3247129    .1515755
         age |   -.149746   .0380902    -3.93   0.000    -.2244016   -.0750905
  ln_pdensity|   .6951984   .1198407     5.80   0.000      .460315    .9300818
------------------------------------------------------------------------------
```

- The marginal (derivative) changes are the same as a unit change, because there are no powers or interactions among the covariates
- A unit increase in the Gini coefficient (on a scale of 0 to 100) is not the most interesting effect
  I calculated the sample standard deviation of the gini coefficients for the Texas counties and use it to scale the change

```
. summarize gini
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        gini |        254    40.33848    3.196163   27.11916   50.04854
```

- Use `margins` to estimate the mean unemployment level in Texas when each county has its observed gini coefficient and when each county has a gini coefficient that is increased by 3.2

```
. margins, at(gini = generate(gini)) at(gini = generate(gini + 3.2))
Predictive margins                              Number of obs     =        254
Model VCE     : OIM
Expression    : Reduced-form mean, predict()
1._at         : gini            = gini
2._at         : gini            = gini + 3.2

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   6.819183   .1470491    46.37   0.000     6.530972    7.107394
          2  |    8.63652   .2267329    38.09   0.000     8.192131    9.080908
------------------------------------------------------------------------------
```

- Use `margins`, `contrast` to estimate the difference in the mean unemployment level when each county has a gini coefficient that is increased by 3.2 and when each county has its observed gini coefficient

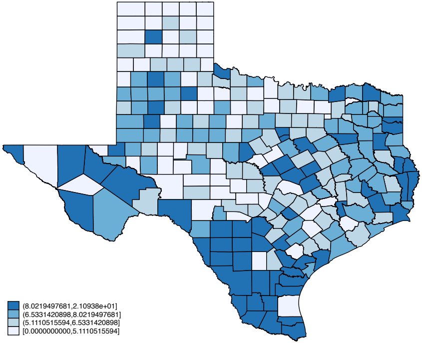
```
. margins, at(gini = generate(gini)) at(gini = generate(gini + 3.2)) ///
>          contrast(at(r) nowald)
Contrasts of predictive margins
Model VCE     : OIM
Expression    : Reduced-form mean, predict()
1._at         : gini             = gini
2._at         : gini             = gini + 3.2

------------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------
         _at |
    (2 vs 1) |   1.817337   .1685768       1.486933    2.147741
------------------------------------------------------------------
```

How I obtained and managed my data

Analysis and shapefile data

```
. use texas_unemp, clear
. grmap unemployment
```

- `texas_unemp` is the analysis data containing the outcome covariate data
- The analysis data is linked to the shapefile data that contains the map information



Step 0: Download shape file

- I downloaded the shape files for US counties from US Census Tiger Line website
  - The file name is `tl_2016_us_county.zip`
  - You can download it from https://www.census.gov/cgi-bin/geo/shapefiles/index.php Specify "2016" for year and "Counties (and equivalent)" for layer type

Step 1a: Extract files

```
. // Downloaded tl_2016_us_county.zip from US Tiger line site
. // unzip downloaded data
. unzipfile tl_2016_us_county
    inflating: tl_2016_us_county.cpg
    inflating: tl_2016_us_county.dbf
    inflating: tl_2016_us_county.prj
    inflating: tl_2016_us_county.shp
    inflating: tl_2016_us_county.shp.ea.iso.xml
    inflating: tl_2016_us_county.shp.iso.xml
    inflating: tl_2016_us_county.shp.xml
    inflating: tl_2016_us_county.shx
successfully unzipped tl_2016_us_county.zip to current directory
total processed:  8
        skipped:  0
      extracted:  8
. //   This creates
. //        inflating: tl_2016_us_county.cpg
. //        inflating: tl_2016_us_county.dbf
. //        inflating: tl_2016_us_county.prj
. //        inflating: tl_2016_us_county.shp
. //        inflating: tl_2016_us_county.shp.ea.iso.xml
. //        inflating: tl_2016_us_county.shp.iso.xml
. //        inflating: tl_2016_us_county.shp.xml
. //        inflating: tl_2016_us_county.shx
```

Step 1b: Extract files

```
. //  We do not need
. //        inflating: tl_2016_us_county.cpg
. //        inflating: tl_2016_us_county.prj
. //        inflating: tl_2016_us_county.shp.ea.iso.xml
. //        inflating: tl_2016_us_county.shp.iso.xml
. //        inflating: tl_2016_us_county.shp.xml
. //        inflating: tl_2016_us_county.shx
. // so I erase them
. erase tl_2016_us_county.cpg
. erase tl_2016_us_county.prj
. erase tl_2016_us_county.shp.ea.iso.xml
. erase tl_2016_us_county.shp.iso.xml
. erase tl_2016_us_county.shp.xml
. erase tl_2016_us_county.shx
```

Step 2: Translate shapefiles

```
. // Translate unziped shapefiles
. // Only
. //        tl_2016_us_county.shp
. //        tl_2016_us_county.dbf
. // are tranlated
. spshape2dta tl_2016_us_county, replace
  (importing .shp file)
  (importing .dbf file)
  (creating _ID spatial-unit id)
  (creating _CX coordinate)
  (creating _CY coordinate)

  file tl_2016_us_county_shp.dta created
  file tl_2016_us_county.dta     created

.
. // No longer need
. //        tl_2016_us_county.shp
. //        tl_2016_us_county.dbf
. // so erase them
. erase tl_2016_us_county.shp
. erase tl_2016_us_county.dbf
```

Step 3a: Describe shapefile

```
. use tl_2016_us_county_shp
. describe
Contains data from tl_2016_us_county_shp.dta
  obs:     7,740,937
 vars:             5                        15 Aug 2018 19:16
 size:   232,655,582
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
_ID             int     %12.0g
_X              double  %10.0g
_Y              double  %10.0g
rec_header      strL    %9s
shape_order     long    %12.0g
--------------------------------------------------------------------------------
Sorted by: _ID
```

Step 3b: list an observation

```
. list _ID _X _Y in 1/10

     +---------------------------------+
     | _ID           _X             _Y |
     |---------------------------------|
  1. |   1            .              . |
  2. |   1   -97.019516      42.004097 |
  3. |   1   -97.019519      42.004933 |
  4. |   1   -97.019527      42.007501 |
  5. |   1   -97.019529      42.009755 |
     |---------------------------------|
  6. |   1   -97.019529      42.009776 |
  7. |   1   -97.019529      42.009939 |
  8. |   1   -97.019529      42.010163 |
  9. |   1   -97.019538      42.013931 |
 10. |   1    -97.01955      42.014546 |
     +---------------------------------+
```

Step 3c: Describe data on places from dbf

```
. use tl_2016_us_county
. describe
Contains data from tl_2016_us_county.dta
  obs:         3,233
 vars:            20                          15 Aug 2018 19:16
 size:       491,416
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
_ID             int     %12.0g                Spatial-unit ID
_CX             double  %10.0g                x-coordinate of area centroid
_CY             double  %10.0g                y-coordinate of area centroid
STATEFP         str2    %9s                   STATEFP
COUNTYFP        str3    %9s                   COUNTYFP
COUNTYNS        str8    %9s                   COUNTYNS
GEOID           str5    %9s                   GEOID
NAME            str21   %21s                  NAME
NAMELSAD        str33   %33s                  NAMELSAD
LSAD            str2    %9s                   LSAD
CLASSFP         str2    %9s                   CLASSFP
MTFCC           str5    %9s                   MTFCC
CSAFP           str3    %9s                   CSAFP
CBSAFP          str5    %9s                   CBSAFP
METDIVFP        str5    %9s                   METDIVFP
FUNCSTAT        str1    %9s                   FUNCSTAT
ALAND           double  %14.0f                ALAND
AWATER          double  %14.0f                AWATER
```

Step 3d: list an observation

```
. list in 1

     +------------------------------------------------------------------------------+
  1. | _ID |         _CX |          _CY | STATEFP | COUNTYFP | COUNTYNS | GEOID |
     |   1 |    -96.7874 |    41.916403 |      31 |      039 | 00835841 | 31039 |
     |------------------------------------------------------------------------------|
     |   NAME |       NAMELSAD | LSAD | CLASSFP | MTFCC | CSAFP | CBSAFP |
     | Cuming |  Cuming County |   06 |      H1 | G4020 |       |        |
     |------------------------------------------------------------------------------|
     | METDIVFP | FUNCSTAT |         ALAND |       AWATER |        INTPTLAT |
     |          |        A |    1477895811 |     10447360 |     +41.9158651 |
     |------------------------------------------------------------------------------|
     |                                INTPTLON                                      |
     |                            -096.7885168                                      |
     +------------------------------------------------------------------------------+
```

Step 3e: Keep sample of interest in shape file

```
. // Keep sample of interest in tl_2016_us_county
. // Only keep data for Texas
. // Texas FIP code is 48
. keep if real(STATEFP) == 48
(2,979 observations deleted)

. generate fips = real(STATEFP + COUNTYFP)

. spcompress, force
  (tl_2016_us_county_shp.dta created with 254 spatial units, 2,979 fewer than
   previously)
  (tl_2016_us_county_shp.dta saved)
  (tl_2016_us_county.dta saved)

. save texas_county, replace
file texas_county.dta saved
```

Step 4: Merge analysis data with shape file data

```
. // Merge utexas data with texas_county shape file data
. use utexas
(S.Messner et al.(2000), U.S southern county homicide rates in 1990)

. merge 1:1 fips using texas_county
(note: variable fips was long, now double to accommodate using data´s values)
    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                               254  (_merge==3)
    -----------------------------------------

. assert _merge == 3

. drop _merge
```

Step 5: spset data

```
. spset , shpfile(tl_2016_us_county_shp) modify
  (creating _ID spatial-unit id)
  (creating _CX coordinate)
  (creating _CY coordinate)
  Sp dataset utexas.dta
                data:  cross sectional
     spatial-unit id:  _ID
         coordinates:  _CX, _CY (planar)
    linked shapefile:  tl_2016_us_county_shp.dta

. save texas_unemp, replace
file texas_unemp.dta saved
```

Ready to go

```
. use texas_unemp, clear
. grmap unemployment
```

Panel data

Basic model

- Consider the model

$$\mathbf{y}_t = \lambda \mathbf{W}\mathbf{y}_t + \mathbf{X}_t\beta + \mathbf{u} + \boldsymbol{\epsilon}_t$$

  where
    - $\mathbf{y}_t$ is the $N \times 1$ vector of outcomes for each $t \in \{1, 2, \ldots, T\}$
    - $\mathbf{X}_t$ is the $N \times k$ matrix of covariates for each $t \in \{1, 2, \ldots, T\}$
    - $\mathbf{u}$ is the $N \times 1$ vector of time-invariant individual level effect
    - $\boldsymbol{\epsilon}_t$ is the $N \times 1$ vector of idiosyncratic errors for each $t \in \{1, 2, \ldots, T\}$
- Fixed effects if $\mathbf{u}$ is correlated with $\mathbf{X}_t$
    - $\mathbf{u}$ are removed prior to estimation
    - All inference is conditional on the unobserved fixed effect
- Random effects if $\mathbf{u}$ is uncorrelated with $\mathbf{X}_t$
    - $\mathbf{u}$ just add a variance component to the model
    - All inference is for the population after the $\mathbf{u}$ are averaged out

Fixed effects

- Recall the model

$$\mathbf{y}_t = \lambda \mathbf{W} \mathbf{y}_t + \mathbf{X}_t \boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}_t$$

- Fixed effects if $\mathbf{u}$ is correlated with $\mathbf{X}_t$
  - Multiply both sides by a matrix that removes the time-invariant component $\mathbf{u}$ prior to estimation

$$\tilde{\mathbf{y}}_t = \lambda \mathbf{W} \tilde{\mathbf{y}}_t + \tilde{\mathbf{X}}_t \boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_t$$

  - After estimating $\boldsymbol{\beta}$, we can predict

$$\mathbf{E}[\check{\mathbf{y}}_t | \mathbf{X}_t] = (\mathbf{I} - \lambda\ \mathbf{W})^{-1} \mathbf{X}_t \boldsymbol{\beta}'$$

    where

$$\check{\mathbf{y}}_t = \mathbf{y}_t - (\mathbf{I} - \lambda \mathbf{W})^{-1} \mathbf{u}$$

  - All inference is conditional on the unobserved fixed effect

Covariate effects

- Solving the model yields

$$\mathbf{E}[\check{y}_{i,t}|\mathbf{X}_t] = s_{i,i}\mathbf{x}_{i,t}\boldsymbol{\beta} + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\mathbf{x}_{j,t}\boldsymbol{\beta}$$

$$\frac{\partial \mathbf{E}[\check{y}_{i,t}|\mathbf{X}_t]}{\partial \mathbf{x}_{t,k}} = s_{i,i}\beta_k + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k$$

- The marginal effect on the mean outcome in place i at time t (minus its fixed effect) of an infinitesimal change in all the observations in time t of covariate k

$$\frac{\partial \mathbf{E}[\check{y}_{i,t}|\mathbf{X}_t]}{\partial \mathbf{x}_{t,k}} = s_{i,i}\beta_k + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k$$

$$\text{Average of total effects (time } t) = 1/n \sum_{i=1}^{n} \left(s_{i,i}\beta_k + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k \right)$$

$$\text{Average of direct effects (time } t) = 1/n \sum_{i=1}^{n} (s_{i,i}\beta_k)$$

$$\text{Mean of indirect effects (time } t) = 1/n \sum_{i=1}^{n} \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k \right)$$

Panel data on Texas unemployment

```
. clear all
. use texas_unemp_60_90, clear
. spmatrix create contiguity C if year==1990
```

FE estimation

```
. spxtregress unemployment c.gini#i.year age ln_pdensity , dvarlag(C) fe nolog
  (1016 observations)
  (1016 observations used)
  (data contain 254 panels (places) )
  (weighting matrix defines 254 places)
Fixed-effects spatial regression                 Number of obs     =      1,016
Group variable: _ID                              Number of groups  =        254
                                                 Obs per group     =          4
                                                 Wald chi2(6)      =    1039.21
                                                 Prob > chi2       =     0.0000
Log likelihood = -1304.4700                      Pseudo R2         =     0.2602
------------------------------------------------------------------------------
unemployment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
unemployment |
 year#c.gini |
        1960 |   .0157626   .0037304     4.23   0.000     .0084511    .0230741
        1970 |          0  (omitted)
        1980 |   .0062974   .0031237     2.02   0.044     .0001749    .0124198
        1990 |   .0632647   .0051919    12.19   0.000     .0530889    .0734406
             |
         age |   .0440119   .0206137     2.14   0.033     .0036098     .084414
 ln_pdensity |    .494616    .227036     2.18   0.029     .0496337    .9395984
-------------+----------------------------------------------------------------
C            |
unemployment |   .2107493    .057986     3.63   0.000     .0970989    .3243998
-------------+----------------------------------------------------------------
```

FE impact

```
. estat impact gini if year == 1960
progress   :100%
Average impacts                              Number of obs     =        254
```

	dy/dx	Delta-Method Std. Err.	z	P>|z|	[95% Conf.	Interval]
direct						
gini	.0158631	.0037411	4.24	0.000	.0085307	.0231954
indirect						
gini	.0034762	.0012581	2.76	0.006	.0010104	.0059419
total						
gini	.0193392	.004453	4.34	0.000	.0106115	.028067

Mundlack controls

- Include panel-level means, also known as Mundlack controls, for relationship between u_i and $\mathbf{x}_{i,t}$

$$u_i = \bar{\mathbf{x}}_i \boldsymbol{\delta} + \xi_i$$

where

$$\bar{\mathbf{x}}_i = 1/T \sum_{t=1}^{T} \mathbf{x}_{i,t}$$

- Allows us to predict the mean of $\mathbf{y}$ having averaged out random effect ξ_i
- Inference is for the population, it is not conditional on u_i fixed effects

Covariate effects

- Solving the model yields

$$\mathbf{E}[y_{i,t}|\mathbf{X}_t] = s_{i,i}\mathbf{x}_{i,t}\boldsymbol{\beta} + s_{i,i}\bar{\mathbf{x}}_i\boldsymbol{\delta} + \sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\mathbf{x}_{j,t}\boldsymbol{\beta} + \sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\bar{\mathbf{x}}_j\boldsymbol{\delta}$$

$$\frac{\partial \mathbf{E}[y_{i,t}|\mathbf{X}_t]}{\partial \mathbf{x}_{t,k}} = s_{i,i}\beta_k + s_{i,i}\delta_k/T + \sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\beta_k + \sum_{\substack{j=1\\ j\neq i}}^{n} s_{i,j}\delta_k/T$$

  - The marginal effect on the mean outcome in place i at time t of an infinitesimal change in all the observations in time t of covariate k

$$\frac{\partial \mathbf{E}[y_{i,t}|\mathbf{X}_t]}{\partial \mathbf{x}_{t,k}} = s_{i,i}\beta_k + s_{i,i}\delta_k/T + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\delta_k/T$$

Average of total effects (time t) =

$$1/n\sum_{i=1}^{n}\left(s_{i,i}\beta_k + s_{i,i}\delta_k/T + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\delta_k/T\right)$$

Average of direct effects (time t) = $1/n\sum_{i=1}^{n}(s_{i,i}\beta_k + s_{i,i}\delta_k/T)$

Mean of indirect effects (time t) =

$$1/n\sum_{i=1}^{n}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\beta_k + \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{i,j}\delta_k/T\right)$$

```
. spxtregress unemployment c.gini#i.year age ln_pdensity  ///
>         gini_m  age_m ln_pdensity_m , dvarlag(C) re nolog
  (1016 observations)
  (1016 observations used)
  (data contain 254 panels (places) )
  (weighting matrix defines 254 places)
Random-effects spatial regression               Number of obs     =      1,016
Group variable: _ID                             Number of groups  =        254
                                                Obs per group     =          4
                                                Wald chi2(10)     =    1264.65
                                                Prob > chi2       =     0.0000
Log likelihood = -1913.0724                     Pseudo R2         =     0.4896

------------------------------------------------------------------------------
unemployment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
unemployment |
 year#c.gini |
       1960  |   .0414289   .0205543     2.02   0.044     .0011432    .0817147
       1970  |   .0220035   .0176232     1.25   0.212    -.0125374    .0565443
       1980  |   .0283204   .0177959     1.59   0.112     -.006559    .0631997
       1990  |   .0864237   .0177105     4.88   0.000     .0517117    .1211356
             |
         age |   .0376273   .0211738     1.78   0.076    -.0038726    .0791273
 ln_pdensity |   .5318764   .2280424     2.33   0.020     .0849216    .9788313
      gini_m |   .2022997   .0337257     6.00   0.000     .1361985    .2684009
       age_m |  -.1611113   .0266186    -6.05   0.000    -.2132828   -.1089398
ln_pdensit~m |  -.1006964   .2359112    -0.43   0.669    -.5630739     .361681
       _cons |  -2.874711   1.177738    -2.44   0.015    -5.183035   -.5663872
-------------+----------------------------------------------------------------
C            |
     ...yment |   .1922244    .043842     4.38   0.000     .1062957    .2781531
```

RE impact

```
. estat impact gini gini_m if year == 1960
progress   : 50% 100%
Average impacts                                 Number of obs     =        254
```

	dy/dx	Delta-Method Std. Err.	z	P>|z|	[95% Conf.	Interval]
direct						
gini	.0416468	.0206672	2.02	0.044	.0011397	.0821538
gini_m	.2033632	.0337917	6.02	0.000	.1371327	.2695938
indirect						
gini	.0081787	.0047121	1.74	0.083	-.0010569	.0174143
gini_m	.0399372	.0113741	3.51	0.000	.0176443	.0622301
total						
gini	.0498255	.0249315	2.00	0.046	.0009607	.0986903
gini_m	.2433005	.0395622	6.15	0.000	.1657601	.3208408

RE impact

```
. margins , at(gini = generate(gini))                                   ///
>         at(gini = generate(gini+1) gini_m= generate(gini_m + 1/4)) ///
>         subpop(if year==1960) contrast(at(r) nowald)

Contrasts of predictive margins
Model VCE     : OIM

Expression    : Reduced-form mean, predict()

1._at         : gini             = gini

2._at         : gini             = gini+1
                gini_m           = gini_m + 1/4

------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
         _at |
   (2 vs 1)  |   .1106506    .021206      .0690876    .1522137
------------------------------------------------------------
```

RE impact

```
. margins , at(gini = generate(gini))                                  ///
>         at(gini = generate(gini+1) gini_m= generate(gini_m + 1)) ///
>         subpop(if year==1960) contrast(at(r) nowald)
Contrasts of predictive margins
Model VCE    : OIM
Expression   : Reduced-form mean, predict()
1._at        : gini              = gini
2._at        : gini              = gini+1
               gini_m            = gini_m + 1

-------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
         _at |
   (2 vs 1)  |    .293126    .0332854      .2278878    .3583641
-------------------------------------------------------------
```

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(2).

Drukker, D. M., P. H. Egger, and I. R. Prucha. 2013a. On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. *Econometric Reviews* 32: 686733.

Drukker, D. M., H. Peng, I. R. Prucha, and R. Raciborski. 2013b. Creating and managing spatial-weigting matrices with the spmat command. *The Stata Journal* 13(2): 242–286.

Drukker, D. M., I. R. Prucha, and R. Raciborski. 2013c. A command for estimating spatial-autoregressive models with spatial-autoregressive disturbances and additional endogenous variables. *The Stata Journal* 13(2): 287–301.

———. 2013d. Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances. *The Stata Journal* 13(2): 221–241.

Kapoor, M., H. H. Kelejian, and I. R. Prucha. 2007. Panel data models with spatially correlated error components. *Journal of Econometrics* 140: 97–130.

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.

———. 1999. A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review* 40(2): 509–533.

———. 2010. Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances. *Journal of Econometrics* 157: 53–67.

Lee, L. F. 2004. Asymptotic distributions of maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72: 1899–1925.

Lee, L. F., and J. Yu. 2010a. Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* 154: 165–185.

———. 2010b. Some recent developments in spatial panel data models. *Regional Science and Urban Economics* 40: 255–271.