Title

Intro 7 — Example from start to finish

Description Remarks and examples Also see

Description

This entry comprises an example from start to finish.

You may also be interested in introductions to other aspects of Sp. Below, we provide links to those other introductions.

Intro 1	A brief introduction to SAR models
Intro 2	The **W** matrix
Intro 3	Preparing data for analysis
Intro 4	Preparing data: Data with shapefiles
Intro 5	Preparing data: Data containing locations (no shapefiles)
Intro 6	Preparing data: Data without shapefiles or locations
Intro 8	The Sp estimation commands

Remarks and examples

stata.com

Remarks are presented under the following headings:

Research plan

We are going to analyze unemployment in counties of Texas. We are going to use `texas_ue.dta`. The data contain unemployment rates and college graduation rates for Texas counties, but they do not include the locations of the counties or a map. The data can be used to fit models with `regress`, but they do not contain the information necessary to fit models with `spregress` that could account for spillover effects.

We will

1. find and download a U.S. counties shapefile,
2. translate the downloaded file to Stata format,
3. merge the translated file with our existing data, and
4. analyze the merged data.

Please keep in mind that this is just an example in a computer software manual. We will model the unemployment rate as a function of college graduation rate only, though we ought to include other explanatory variables. We analyze data for Texas only, though we should use the entire United States. We will draw conclusions that are unjustified, and we will not qualify them appropriately. We will, however, show you how to use `spregress` and interpret its output.

Finding and preparing data

We first find and download an appropriate shapefile from the web. Then, we will prepare it as described in [SP] **Intro 4**.

Finding a shapefile for Texas counties

We looked for a county shapefile for Texas but could not find one. We did find shapefiles for the entire United States, however. We used our browser to search for "shapefile U.S. counties census". From the results, we selected *TIGER/Line Shapefile, 2016, nation, U.S., Current County and Equivalent National Shapefile*. On the resulting page, we clicked to download the **Shapefile Zip File** from the **Downloads & Resources** section. File `tl_2016_us_county.zip` was downloaded to the `Downloads` directory on our computer.

Creating the Stata-format shapefile

We found a standard-format shapefile, `tl_2016_us_county.zip`. We now follow the instructions in [SP] **Intro 4** to create a Stata-format shapefile. Here is the result:

```
. // ----------------------------------------------------------
. // [SP] intro 4, step 2: Translate the shapefile
.
. copy ~/Downloads/tl_2016_us_county.zip .
. unzipfile tl_2016_us_county.zip
    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
```

```
. spshape2dta tl_2016_us_county
  (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

.
. // ----------------------------------------------------------
. // [SP] intro 4, step 3: Look at the data
.
. use tl_2016_us_county, clear

. describe

Contains data from tl_2016_us_county.dta
 Observations:          3,233
    Variables:             20                  9 Feb 2023 12:44
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    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
INTPTLAT        str11   %11s                  INTPTLAT
INTPTLON        str12   %12s                  INTPTLON
-------------------------------------------------------------------------------
Sorted by: _ID

. list in 1/2

    +-----------------------------------------------------------------------------+
 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                                  |
    +-----------------------------------------------------------------------------+
```

```
  2.  | _ID           _CX         _CY   STATEFP   COUNTYFP   COUNTYNS   GEOID |
      |   2    -123.43347   46.291134        53        069   01513275   53069 |
      |---------------------------------------------------------------------|
      |      NAME              NAMELSAD   LSAD   CLASSFP   MTFCC   CSAFP   CBSAFP |
      | Wahkiakum      Wahkiakum County     06        H1   G4020                  |
      |---------------------------------------------------------------------|
      | METDIVFP   FUNCSTAT       ALAND     AWATER      INTPTLAT |
      |                   A   680956787   61588406   +46.2946377 |
      |---------------------------------------------------------------------|
      |                          INTPTLON                                    |
      |                      -123.4244583                                    |

.
. // ----------------------------------------------------------
. // [SP] intro 4, step 4: Create standard ID variable
.
. generate long fips = real(STATEFP + COUNTYFP)
. bysort fips: assert _N==1
. assert fips != .

.
. // ----------------------------------------------------------
. // [SP] intro 4, step 5: Tell Sp to use standard ID variable
.
. spset fips, modify replace
  (_shp.dta file saved)
  (data in memory saved)
      Sp dataset: tl_2016_us_county.dta
Linked shapefile: tl_2016_us_county_shp.dta
            Data: Cross sectional
 Spatial-unit ID: _ID (equal to fips)
     Coordinates: _CX, _CY (planar)

.
. // ----------------------------------------------------------
. // [SP] intro 4, step 6: Set coordinate units
.
. spset, modify coordsys(latlong, miles)
      Sp dataset: tl_2016_us_county.dta
Linked shapefile: tl_2016_us_county_shp.dta
            Data: Cross sectional
 Spatial-unit ID: _ID (equal to fips)
     Coordinates: _CY, _CX (latitude-and-longitude, miles)

. save, replace
file tl_2016_us_county.dta saved
. // ----------------------------------------------------------
```

Merging our data with the Stata-format shapefile

Recall that we are going to use `texas_ue.dta` containing unemployment rates and college graduation rates for Texas counties. We follow the instructions in [SP] **Intro 4**, *Step 7a* to merge our existing data with the Stata-format shapefile.

```
. copy https://www.stata-press.com/data/r18/texas_ue.dta .
. use texas_ue, clear
. describe
Contains data from texas_ue.dta
 Observations:           254
    Variables:             4                  10 Feb 2023 12:36
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
fips            float   %9.0g                 FIPS
college         float   %9.0g               * Percent college degree
income          long    %12.0g                Median household income
unemployment    float   %9.0g                 Unemployment rate
                                            * indicated variables have notes
-------------------------------------------------------------------------------
Sorted by: fips
. merge 1:1 fips using tl_2016_us_county
(variable fips was float, now double to accommodate using data's values)
    Result                      Number of obs
    -----------------------------------------
    Not matched                         2,979
        from master                         0  (_merge==1)
        from using                      2,979  (_merge==2)
    Matched                               254  (_merge==3)
    -----------------------------------------
. keep if _merge==3
(2,979 observations deleted)
. drop _merge
```

At this point, we type `describe` again and discover that `texas_ue.dta` has lots of unnecessary, leftover variables from `tl_2016_us_county.dta`, so we drop them. There is another variable that we rather like—the names of the counties—and we rename it.

```
. rename NAME countyname
. drop STATEFP COUNTYFP COUNTYNS GEOID
. drop NAMELSAD LSAD CLASSFP MTFCC CSAFP
. drop CBSAFP METDIVFP FUNCSTAT
. drop ALAND AWATER INTPTLAT INTPTLON
. save, replace
file texas_ue.dta saved
```

Analyzing texas_ue.dta

File `texas_ue.dta` is our updated analysis dataset that can be used with Sp commands.

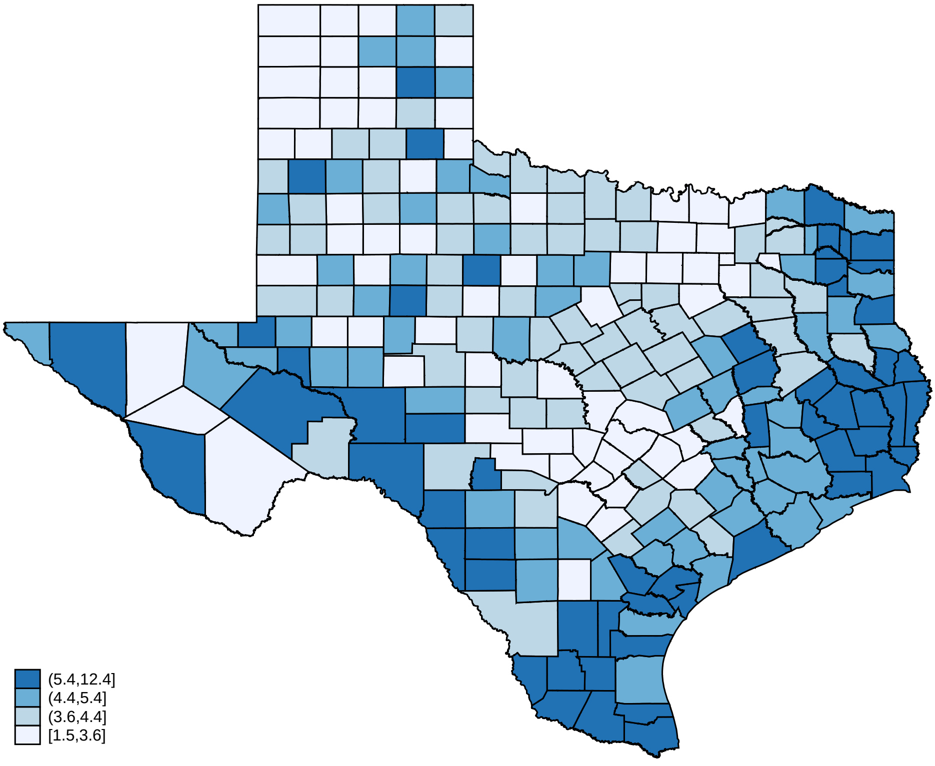
```
. describe

Contains data from texas_ue.dta
 Observations:           254
    Variables:             8                 24 Mar 2023 21:54
                                             (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
fips            double  %9.0g                 FIPS
college         float   %9.0g               * Percent college degree
income          long    %12.0g                Median household income
unemployment    float   %9.0g                 Unemployment rate
_ID             long    %12.0g                Spatial-unit ID
_CX             double  %10.0g                x-coordinate of area centroid
_CY             double  %10.0g                y-coordinate of area centroid
countyname      str21   %21s                  NAME
                                            * indicated variables have notes
-------------------------------------------------------------------------------
Sorted by:
```

Our example uses the unemployment rate. It varies between 1.5% and 12.4% across the counties of Texas:

```
. summarize unemployment
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
unemployment |        254    4.731102    1.716514        1.5       12.4
```

Because `texas_ue.dta` has been `spset` and has a shapefile, we can draw choropleth maps, such as this one of the unemployment rate:

```
. grmap unemployment
```



Unemployment appears to be clustered, which suggests that there are spillover effects between counties.

Testing whether ordinary regression is adequate

These data are suitable for both spatial and nonspatial analysis. (Spatial data always are.) We will fit a linear regression of the unemployment rate on the college graduation rate, mostly for illustrative purposes. After fitting the linear regression, we will use an Sp command to determine whether the residuals of the model are spatially correlated, and we find that they are.

Here is the regression:

```
. regress unemployment college

      Source |       SS           df       MS      Number of obs   =       254
-------------+----------------------------------   F(1, 252)       =     57.92
       Model |  139.314746         1  139.314746   Prob > F        =    0.0000
    Residual |  606.129539       252  2.40527595   R-squared       =    0.1869
-------------+----------------------------------   Adj R-squared   =    0.1837
       Total |  745.444285       253   2.9464201   Root MSE        =    1.5509

------------------------------------------------------------------------------
unemployment | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     college |  -.1008791   .0132552    -7.61   0.000    -.1269842   -.0747741
       _cons |   6.542796   .2571722    25.44   0.000     6.036316    7.049277
------------------------------------------------------------------------------
```

The results of this oversimplified model indicate that the college graduation rate reduces unemployment markedly.

Are we done? If the residuals show no signs of being spatially clustered, then we are. We can perform a statistical test.

Sp provides the Moran test for determining whether the residuals of a model fit by `regress` are correlated with nearby residuals. To use it, we must define "nearby". We do that by defining a spatial weighting matrix, which is created by the `spmatrix` command. We will define a contiguity matrix.

```
. spmatrix create contiguity W
```

This contiguity matrix sets "nearby" to mean "shares a border".

`spmatrix` can create other types of weighting matrices. It even allows you to create custom matrices or to import matrices. See [SP] **spmatrix**.

We can now run the Moran test.

```
. estat moran, errorlag(W)
Moran test for spatial dependence
         H0: Error terms are i.i.d.
         Errorlags:  W

         chi2(1)      =    94.06
         Prob > chi2  =   0.0000
```

The test reports that we can reject that the residuals from the model above are independent and identically distributed (i.i.d.). In particular, the test considered the alternative hypothesis that residuals are correlated with nearby residuals as defined by $\mathbf{W}$.

spregress can reproduce regress results

`spregress` is the spatial autoregression command. `spregress` fits models in which the observations are not independent, as defined by the $\mathbf{W}$ weighting matrix.

Above, we fit a model under the assumption that the counties are independent. We used `regress`, Stata's ordinary linear regression command. We typed

```
. regress unemployment college
```

We could have fit the same model and obtained the same results by using `spregress`. We would have typed

```
. spregress unemployment college, gs2sls
```

or

```
. spregress unemployment college, ml
```

`spregress` is seldom used for fitting models without spatial lags or autocorrelated errors, but when it is, it reports the same linear regression results that `regress` reports, although there are some differences. Standard errors are slightly different, and `spregress` reports Z and χ^2 statistics instead of t and F statistics. `spregress` does not include the finite-sample adjustments that `regress` does because it does not expect to be used in situations where those adjustments would be appropriate.

Fitting models with a spatial lag of the dependent variable

We will use `spregress` to fit the same model we fit using `regress` but with the addition of a spatial lag of unemployment. The model we fit will be

$$\mathbf{y}_{\mathrm{ue}} = \beta_0 + \beta_1 \mathbf{x}_{\mathrm{cr}} + \beta_2 \mathbf{W} \mathbf{y}_{\mathrm{ue}} + \boldsymbol{\epsilon}$$

$\mathbf{y}_{\mathrm{ue}}$ is the unemployment rate corresponding to variable `unemployment` in our data. $\mathbf{x}_{\mathrm{cr}}$ is the college graduation rate corresponding to variable `college`.

The model we fit will include the term $\beta_2 \mathbf{W} \mathbf{y}_{\mathrm{ue}}$, meaning that we will assume the unemployment rate spills over from nearby counties. There is a real logic to such a model. One would expect workers in high unemployment counties to seek employment nearby.

`spregress` provides two ways of fitting models: generalized spatial two-stage least squares (`gs2sls`) and maximum likelihood (`ml`). To fit the above model, we could type

```
. spregress unemployment college, gs2sls dvarlag(W)
```

or

```
. spregress unemployment college, ml dvarlag(W)
```

`spregress, ml` is statistically more efficient than `gs2sls` when the errors are normally distributed. Efficiency is desirable, so we should use `ml`, right? That same property said differently is that `gs2sls` is robust to violations of normality. Robustness is desirable, too. So now the choice between them hinges on whether we believe the normality assumption. That said, `ml` will provide standard errors that are also robust to violations of normality if we specify its `vce(robust)` option. Finally, `ml` takes longer to run, and that computation time increases as the number of observations increases. We will use `gs2sls`.

```
. spregress unemployment college, gs2sls dvarlag(W)
  (254 observations)
  (254 observations (places) used)
  (weighting matrix defines 254 places)

Spatial autoregressive model                   Number of obs =     254
GS2SLS estimates                               Wald chi2(2)  =   67.66
                                               Prob > chi2   = 0.0000
                                               Pseudo R2     = 0.1453

unemployment   Coefficient  Std. err.      z    P>|z|     [95% conf. interval]

unemployment
     college    -.0939834   .0131033    -7.17   0.000    -.1196653   -.0683015
       _cons     5.607379   .5033813    11.14   0.000     4.620769    6.593988

W
unemployment     .2007728   .0942205     2.13   0.033      .016104    .3854415

Wald test of spatial terms:          chi2(1) = 4.54       Prob > chi2 = 0.0331
```

Results for β_0 and β_1 are similar to those reported by `regress`, but that is a fluke of this example. Usually, when spillover effects are significant, other parameters change. Meanwhile, we find that β_2 (which multiplies $\mathbf{W}\mathbf{y}_{\mathrm{ue}}$) is significant, but it is not sharply estimated. The 95% confidence interval places β_2 in the range $[0.02, 0.39]$.

Interpreting models with a spatial lag of the dependent variable

You might be tempted to think of β_1 as the direct effect of education and β_2 as the spillover effect, but they are not. They are ingredients into a recursive calculation of those effects. The model we fit is

$$\mathbf{y}_{\mathrm{ue}} = \beta_0 + \beta_1\mathbf{x}_{\mathrm{cr}} + \beta_2\mathbf{W}\mathbf{y}_{\mathrm{ue}} + \boldsymbol{\epsilon}$$

If $\mathbf{x}_{\mathrm{cr}}$ increases, that reduces $\mathbf{y}_{\mathrm{ue}}$ by β_1, and that reduction in $\mathbf{y}_{\mathrm{ue}}$ spills over to produce a further reduction in $\mathbf{y}_{\mathrm{ue}}$ of $\beta_2\mathbf{W}$, and that reduction spills over to produce yet another reduction in $\mathbf{y}_{\mathrm{ue}}$, and so on.

`estat impact` reports the average effects from the recursive process.

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

                         Delta-Method
                   dy/dx   std. err.      z    P>|z|     [95% conf. interval]

direct
     college   -.0945245   .0130576    -7.24   0.000     -.120117   -.0689321

indirect
     college   -.0195459    .010691    -1.83   0.068       -.0405    .0014081

total
     college   -.1140705   .0171995    -6.63   0.000    -.1477808   -.0803602
```

In these data, both the unemployment and the graduation rates are measured in percentage points. A change of 1 is a change of 1 percentage point. The table above reports derivatives, but we can be forgiven for interpreting the results as if they were for a one-unit change. Everybody does it, and sometimes it is even justifiable, for example, if the model is linear in the variables as this one is. Even if the model were nonlinear, it would be a tolerable approximation to the truth as long as a one-unit change were small.

The table reports average changes for a 1-percentage-point increase in the college graduation rate. The direct effect is the effect of the change within the county, ignoring spillover effects. The own-county direct effect is to reduce the unemployment rate by 0.09 percentage points.

The indirect effect is the spillover effect. A 1-percentage-point increase in the college graduation rate reduces unemployment, and that reduction spills over to further reduce unemployment. The result is a 0.02 reduction in unemployment.

The total effect is the sum of the direct and indirect effects, which is $-0.09 + -0.02 = -0.11$.

You must use `estat impact` to interpret effects. Do not try to judge them from the coefficients that `spregress` reports because they can mislead you. For instance, if we multiplied variable `unemployment` by 100, that would not substantively change anything about the model, yet the effect on the coefficients that `spregress` estimates is surprising.

Summary of `spregress` results

Regression of `unemployment` and `100*unemployment` on `college` and `W*unemployment`

	unemployment	100*unemployment
college	−0.094	−9.4
W*unemployment	0.201	0.201

Notes: Column 1 from `spregress` output above.
Column 2 from:

```
generate ue100 = 100*unemployment
spregress unemployment college, gs2sls dvarlag(W)
```

The effect of the change in units is to multiply the coefficient on `college` (β_1) by 100 just as you would expect. Yet β_2, the coefficient on $\mathbf{W}\mathbf{y}_{\rm ue}$, is unchanged! Comparing these two models, you might mislead yourself into thinking that the ratio of the indirect-to-direct effects is smaller in the second model, but it is not. `estat impact` continues to report the same results as it did previously, multiplied by 100:

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

------------------------------------------------------------------------------
             |            Delta-Method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
direct       |
     college |  -9.452455    1.30576    -7.24   0.000     -12.0117   -6.893213
-------------+----------------------------------------------------------------
indirect     |
     college |  -1.954593   1.069105    -1.83   0.068        -4.05    .1408134
-------------+----------------------------------------------------------------
total        |
     college |  -11.40705   1.719946    -6.63   0.000    -14.77808   -8.036016
------------------------------------------------------------------------------
```

Fitting models with a spatial lag of independent variables

We fit a model above with a spatial lag of the dependent variable:

$$\mathbf{y}_{\rm ue} = \beta_0 + \beta_1\mathbf{x}_{\rm cr} + \beta_2\mathbf{W}\mathbf{y}_{\rm ue} + \boldsymbol{\epsilon}$$

We could instead fit a model with a spatial lag of the independent variable:

$$\mathbf{y}_{\rm ue} = \beta_0 + \beta_1\mathbf{x}_{\rm cr} + \beta_2\mathbf{W}\mathbf{x}_{\rm cr} + \boldsymbol{\epsilon}$$

We do that by typing

```
. spregress unemployment college, gs2sls ivarlag(W:college)
  (254 observations)
  (254 observations (places) used)
  (weighting matrix defines 254 places)
Spatial autoregressive model                         Number of obs =    254
GS2SLS estimates                                     Wald chi2(2)  =  81.13
                                                     Prob > chi2   = 0.0000
                                                     Pseudo R2     = 0.2421
```

unemployment	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
unemployment						
college	-.077997	.0138127	-5.65	0.000	-.1050695	-.0509245
_cons	7.424453	.3212299	23.11	0.000	6.794854	8.054053
W						
college	-.0823959	.0191586	-4.30	0.000	-.1199461	-.0448458

```
Wald test of spatial terms:          chi2(1) = 18.50      Prob > chi2 = 0.0000
```

Interpreting models with a spatial lag of the independent variables

Just as with lags of the dependent variable, the easy way to obtain the direct and indirect effects of independent variables is to use `estat impact`.

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

	dy/dx	Delta-Method std. err.	z	P>|z|	[95% conf.	interval]
direct						
college	-.077997	.0138127	-5.65	0.000	-.1050695	-.0509245
indirect						
college	-.0715273	.0166314	-4.30	0.000	-.1041243	-.0389303
total						
college	-.1495243	.0170417	-8.77	0.000	-.1829255	-.1161231

The table reports that the own-county direct effect of a 1-percentage-point increase in the college graduation rate is to reduce unemployment by 0.078 percentage points.

The across-county spillover effect of a 1-percentage-point increase in the college graduation rate is to reduce unemployment by 0.072 percentage points on average.

For those curious how the results were calculated, here are the details.

- The direct effect of college graduation rate is $\beta_1 \mathbf{x}_{\rm cr}$.
- The indirect effect of college graduation rate is $\beta_2 \mathbf{W} \mathbf{x}_{\rm cr}$.
- The direct effect of increasing $\mathbf{x}_{\rm cr}$ by 1 in all counties is

  $$\Delta \mathbf{y}_{\rm ue} = \beta_1(\mathbf{x}_{\rm cr} + \mathbf{1}) - \beta_1 \mathbf{x}_{\rm cr} = \beta_1 \mathbf{1}$$

  where $\mathbf{1}$ is an $N \times 1$ vector of 1s.
- The direct effect is that $\mathbf{y}_{\rm ue}$ increases by β_1 in each county.
- The indirect effect follows the same logic:

  $$\Delta \mathbf{y}_{\rm ue} = \beta_2 \mathbf{W}(\mathbf{x}_{\rm cr} + \mathbf{1}) - \beta_2 \mathbf{W} \mathbf{x}_{\rm cr} = \beta_2 \mathbf{W} \mathbf{1}$$

This result states that $\mathbf{y}_{\rm ue}$ increases by $(\beta_2 \mathbf{W} \mathbf{1})_i$ in county i. For different counties, there are different effects because each county is affected by its own neighbors. The average effect across counties is the average of $\beta_2 \mathbf{W} \mathbf{1}$.

Fitting models with spatially autoregressive errors

We have fit models with a spatial lag of the dependent variable and with a spatial lag of the independent variable.

$$\mathbf{y}_{\rm ue} = \beta_0 + \beta_1 \mathbf{x}_{\rm cr} + \beta_2 \mathbf{W} \mathbf{y}_{\rm ue} + \boldsymbol{\epsilon}$$
$$\mathbf{y}_{\rm ue} = \beta_0 + \beta_1 \mathbf{x}_{\rm cr} + \beta_2 \mathbf{W} \mathbf{x}_{\rm cr} + \boldsymbol{\epsilon}$$

We could instead fit a model with a spatial lag of the error:

$$\mathbf{y}_{\rm ue} = \beta_0 + \beta_1 \mathbf{x}_{\rm cr} + (\mathbf{I} - \rho \mathbf{W})^{-1} \boldsymbol{\epsilon}$$

We do that by typing

```
. spregress unemployment college, gs2sls errorlag(W)
  (254 observations)
  (254 observations (places) used)
  (weighting matrix defines 254 places)
Estimating rho using 2SLS residuals:
Initial:      GMM criterion =  .71251706
Alternative:  GMM criterion =  .04381608
Rescale:      GMM criterion =  .02453154
Iteration 0:  GMM criterion =  .02453154
Iteration 1:  GMM criterion =  .00420723
Iteration 2:  GMM criterion =   .0002217
Iteration 3:  GMM criterion =  .00021298
Iteration 4:  GMM criterion =  .00021298
Estimating rho using GS2SLS residuals:
Iteration 0:  GMM criterion =  .00566696
Iteration 1:  GMM criterion =  .00486118
Iteration 2:  GMM criterion =  .00486066
Iteration 3:  GMM criterion =  .00486066
Spatial autoregressive model                        Number of obs =     254
GS2SLS estimates                                    Wald chi2(1)  =   37.76
                                                    Prob > chi2   = 0.0000
                                                    Pseudo R2     = 0.1869

------------------------------------------------------------------------------
unemployment | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
unemployment |
     college |  -.0759125   .0123532    -6.15   0.000    -.1001243   -.0517008
       _cons |   6.292997   .2968272    21.20   0.000     5.711227    6.874768
-------------+----------------------------------------------------------------
W            |
e.unemploy~t |   .7697395   .0690499    11.15   0.000     .6344043    .9050748
------------------------------------------------------------------------------
Wald test of spatial terms:          chi2(1) = 124.27     Prob > chi2 = 0.0000
```

The estimated value of the spatial autocorrelation parameter ρ is presented on the line above the Wald test: $\widehat{\rho} = 0.77$. It is estimated to be large and significant.

ρ is called the autocorrelation parameter because it is not a correlation coefficient, although it does share some characteristics with correlation coefficients. It is theoretically bounded by -1 and 1, and $\rho = 0$ means that the autocorrelation is 0.

`estat impact` does not report ρ:

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

	dy/dx	Delta-Method std. err.	z	P>|z|	[95% conf.	interval]
direct						
college	-.0759125	.0123532	-6.15	0.000	-.1001243	-.0517008
indirect						
college	0	(omitted)				
total						
college	-.0759125	.0123532	-6.15	0.000	-.1001243	-.0517008

The above output is an example of what `estat impact` produces when there are no lagged dependent or independent variables. There are no spillover effects. Spatially correlated errors do not induce spillover effects in the covariates.

Models can have all three kinds of spatial lag terms

We have shown models with each type of spatial lag term, but models can have more than one. Use `estat impact` to estimate the effects of covariates when you have lagged variables, whether dependent, independent, or both. If you include spatially correlated errors, check the size and significance of the estimated ρ.

Also see

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

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

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

[SP] **spset** — Declare data to be Sp spatial data

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