Applied Spatial Econometrics

Bernard Fingleton
University of Cambridge
UK

Topics to discuss

- Regression and spatial dependence
  - Residual Spatial autocorrelation
- Modelling spatial dependence
  - Spatial lag model, Spatial error model, Spatial Durbin model
- Estimation
  - Two stage least squares (2SLS)
- Software
- ‘how to do spatial econometrics’ in Excel (is it possible?)

The emergence of spatial econometrics?

- Spatial economics now widely recognised in the economics/econometrics mainstream
- Krugman's Nobel prize for work on economic geography
- Importance of network economics (eg Royal Economic Society Easter 2009 School , on 'Auctions and Networks')
- LSE ESRC Centre for Spatial Economics
- Increasing policy relevance : World Bank (2008), *World Development Report 2009,* World Bank, Washington.
- Importantly, much insight can be gained by using spatial econometric tools in addition to more standard time series methods
- Time series methods and spatial econometrics come together in the analysis of spatial panels

What is spatial econometrics?

- the theory and methodology appropriate to the analysis of spatial series relating to the economy
- spatial series means each variable is distributed not in time as in conventional, mainstream econometrics, but in space.

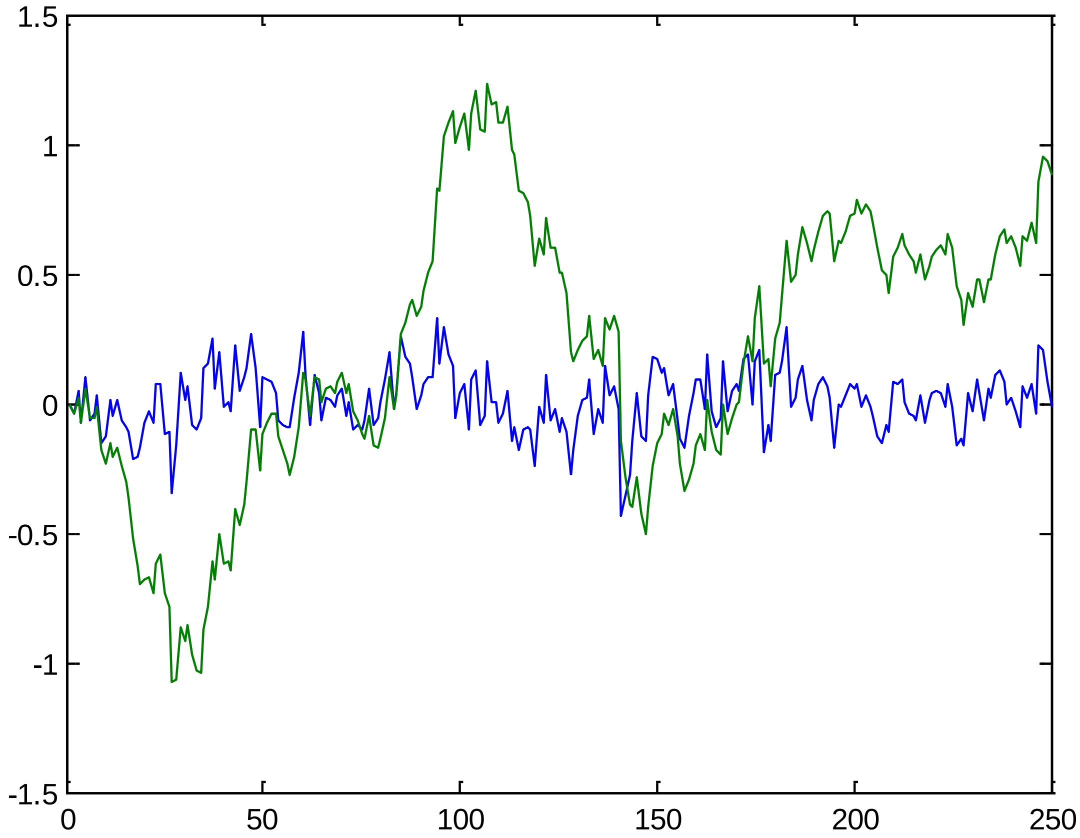
Spatial versus time series

- DGP for time series

$$
\begin{aligned}
&y(t) = \alpha y(t-1) + \varepsilon(t) \\
&\varepsilon(1) = y(1) = 0 \\
&\varepsilon \sim iid(0, \sigma^2) \\
&t = 2 \ldots T
\end{aligned}
$$

Spatial versus time series

- DGP for time series



Spatial versus time series

- DGP for time series

$$y = \alpha W y + \varepsilon$$

y is a T x 1 vector

α is a scalar parameter that is estimated

ε is an T x 1 vector of disturbances

DGP for time series

$$y = \alpha W y + \varepsilon$$

W is a TxT matrix with 1s on the minor diagonal, thus for T = 10

$$W = \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{matrix}$$

The 1s indicate location pairs that are close to each other in time

DGP for time series

$$y = \alpha W y + \varepsilon$$

Provided Wy and ε are contemporaneously independent we can estimate α by OLS and get consistent estimates, although there is small sample bias.

DGP for spatial series

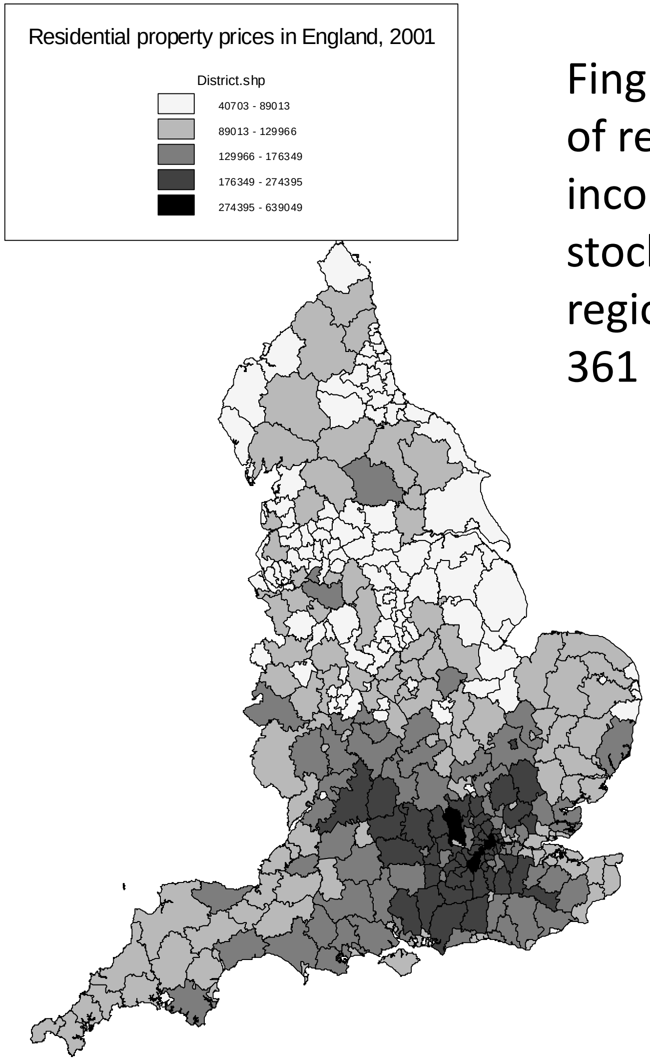
In spatial econometrics, we have an N x N *W* matrix
N is the number of places.

$$
W = \begin{array}{cccccccccc}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0
\end{array}
$$

N= 353
a portion of the W matrix for Luton(1), Mid Bedfordshire(2), Bedford(3) , South Bedfordshire(4), Bracknell Forest(5), Reading(6), Slough(7), West Berkshire(8), Windsor and Maidenhead(9), Wokingham(10)

The 1s indicate location pairs that are close to each other in space

DGP for spatial series



Fingleton B (2006) 'A cross-sectional analysis of residential property prices: the effects of income, commuting, schooling, the housing stock and spatial interaction in the English regions' *Papers in Regional Science* 85 339-361

N= 353

We refer to these small areas
As UALADs

DGP for spatial series

$$y = \rho W y + \varepsilon$$

y is an N x 1 vector

ρ is a scalar parameter that is estimated

ε is an N x 1 vector of disturbances

DGP for spatial series

$$y = \rho Wy + \varepsilon$$

- This is an almost identical set-up to the time series case
  And one might think that it can also be consistently estimated by OLS

- But now there is one big difference

- we cannot estimate the spatial autoregression by OLS
  and obtain consistent estimates of ρ .

- Reason - Wy and ε are not independent.

- Wy determines y but is also determined by y.

But more about this later……

Regression and spatial dependence

- Typically in economics we working with regression models, thus

$$y_t = \sum_k x_{tk} \beta_k + \varepsilon_t$$

- But in spatial economics typically the analysis is cross-sectional, so that

$$y_i = \sum_k x_{ik} \beta_k + \varepsilon_i$$

Regression and spatial dependence

$$y_i = \sum_k x_{ik} \beta_k + \varepsilon_i$$

y_i = Observed value of dependent variable y at location i (i = 1,…,N)

x_{ik} = Observation on explanatory variable x_k at location i, with k = 1,…,K

β_k = regression coefficient for variable x_k

ε_i = random error term or disturbance term at location i

Let us assume as in the classic regression model that the errors ε_i simply represent unmodelled effects that appears to be random. We therefore commence by assuming that $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, E(\varepsilon_i, \varepsilon_j) = 0$ for all i,j . The assumption is that the errors are identically and independently distributed. For the purposes of inference we might specify the error as a normal distribution.

Regression and spatial dependence

- Writing our model in matrix terms gives

$$y = X\beta + \varepsilon$$

y is an N x 1 vector

X is an N x k matrix

β is a k x 1 vector

ε is an N x 1 vector

$$E(\varepsilon) = 0, E(\varepsilon\varepsilon') = \sigma^2 I$$

- And spatial dependence manifests itself as spatially autocorrelated residuals

$$\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta}$$

Residual Spatial autocorrelation

- This term is analogous to autocorrelation in time series, which is when the residuals at points that are close to each other in time/space are not independent.
  - For instance they may be more similar than expected (positive autocorrelation) for some reason.
- suggesting that something is wrong with the model specification that is assuming they are independent.
  - For example the errors/disturbances/residuals may contain the effects of omitted effects that vary systematically across space.

Moran's *I*

- Based on *W* matrix
  - A spatial weights matrix is an N x N with non-zero elements in each row *i* for those columns *j* that are in some way neighbours of location *i*
  - The notion of neighbour is a very general one, it may mean that they are close together in terms of miles or driving time, or it may be distance in some more abstract economic space or social space that is not really connected to geographical distance.
  - The simplest form of distance might be contiguity, with W_{ij}= 1 if locations i and j are contiguous, and W_{ij} = 0 otherwise.
  - Usually (but not necessarily) *W* is standardised so that all the values in row *i* are divided by the sum of the row *i* values.

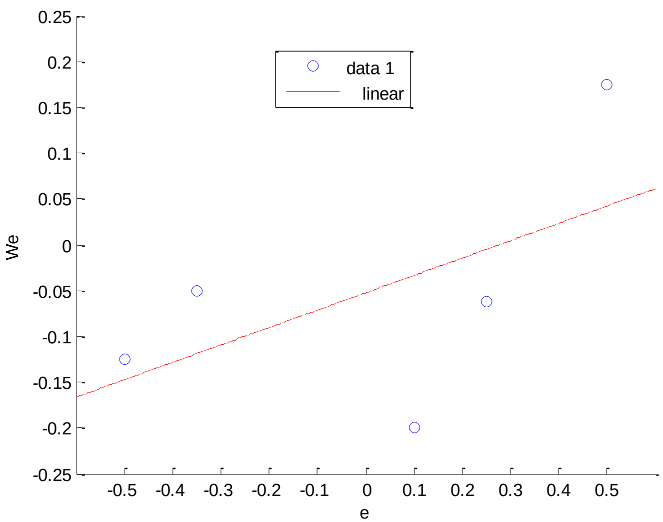
Calculating Moran's *I*

think of Moran's I as approximately the correlation between the two vectors $W\hat{\varepsilon}$ and $\hat{\varepsilon}$. We can show this for a 5 location analysis in graphical form, known as a Moran scatterplot.

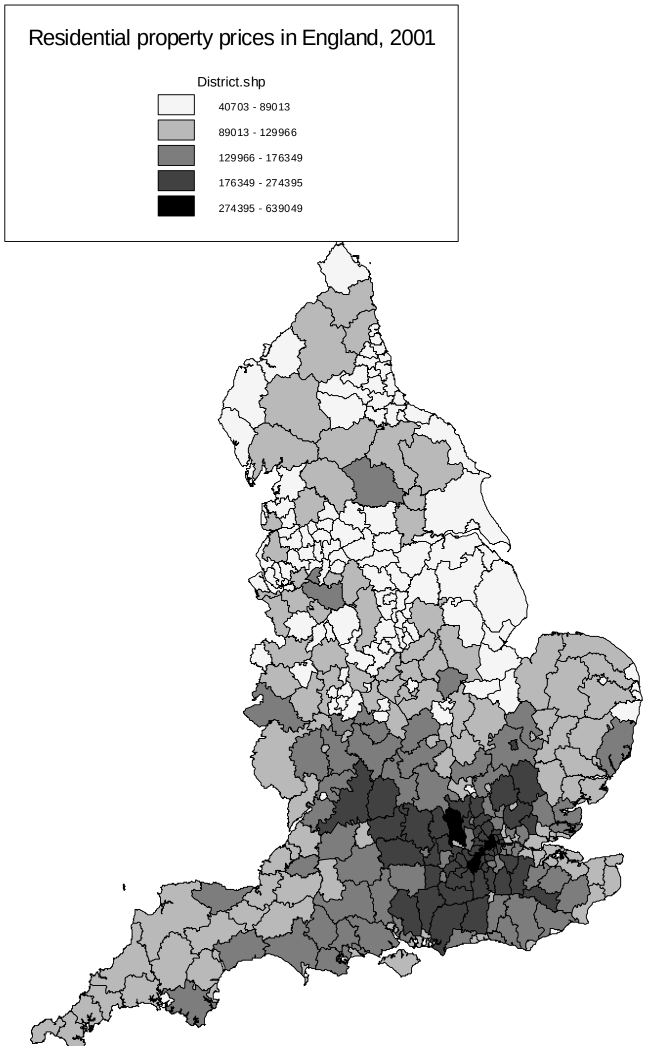
$$W = \begin{matrix} 0 & 0.5000 & 0.5000 & 0 & 0 \\ 0.3330 & 0 & 0.3330 & 0.3330 & 0 \\ 0.3330 & 0.3330 & 0 & 0.3330 & 0 \\ 0.2500 & 0.2500 & 0.2500 & 0 & 0.2500 \\ 0 & 0 & 0.5000 & 0.5000 & 0 \end{matrix} \qquad \hat{\varepsilon} = \begin{matrix} -0.5000 \\ -0.3500 \\ 0.1000 \\ 0.2500 \\ 0.5000 \end{matrix}$$

Hence -0.1250= 0.5 x -0.35 + 0.5 x 0.1.

$$W\hat{\varepsilon} = \begin{matrix} -0.1250 \\ -0.0500 \\ -0.1998 \\ -0.0625 \\ 0.1750 \end{matrix}$$



Average House prices in local authority areas in England (UALADs)



N= 353

Calculating Moran's I in practice

- Let us look at our map of house prices.
- Can we build a model explaining this variation?
- Do we have spatially autocorrelated residuals?
  - The presence of spatial autocorrelation would suggest there is some specification error,
    - either omitted spatially autocorrelated variable
    - residual heterogeneity
    - or a spatial autoregressive error process

Calculating Moran's I in practice

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

y = mean residential property price in each of N local authority areas

X_1 = 1, the constant, an N x 1 vector of 1s

X_2 = total income in each local authority area

X_3 = income earned within commuting distance of each local authority area

X_4 = local schooling quality in each local authority area

X_5 = stock of properties in each local authority area

$$y = X\beta + \varepsilon$$

X is a N x k matrix

β is a k x1 vector

$$\hat{\varepsilon} = y - X\hat{\beta}$$

Dependent variable *y*		
	estimate	t ratio
Constant *(X_1)*	-571.874	-6.47
Local income *(X_2)*	864.0059	10.02
Within-commuting-distance income *(X_3)*	57.7055	14.08
Schooling quality *(X_4)*	175802.9235	7.74
Number of households *(X_5)*	-0.7112	-6.46
R^2 adjusted	0.567	
Standard Error	42.113	
Moran's I	0.39369	11.29
Degrees of freedom	348	

the value for Moran's I is 11.29 standard errors above expectation. Expectation is the expected value of I under the null hypothesis of no residual autocorrelation. It is clear that there is very significant residual autocorrelation.

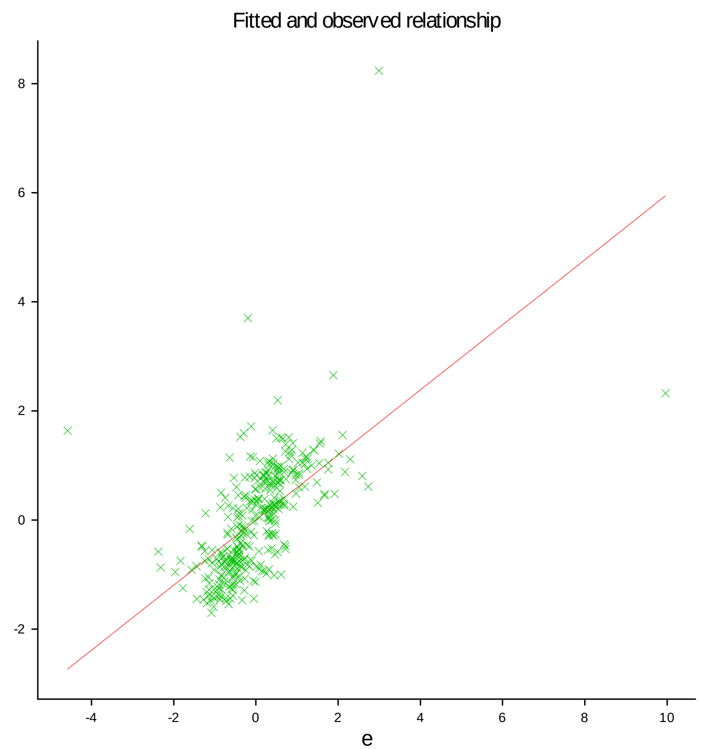
Calculating Moran's I in practice

- What is W?

$$W^* = \frac{1}{d_{ij}^2} \qquad W_{ij} = \frac{W_{ij}^*}{\sum_j W_{ij}^*}$$

Moran scatterplot

$$W\hat{\varepsilon} \text{ versus } \hat{\varepsilon}$$



Calculating Moran's I in practice

The classic formula for Moran's *I* is

$$I = \frac{\hat{\varepsilon}' W \hat{\varepsilon} / S_0}{\hat{\varepsilon}' \hat{\varepsilon} / N}$$

$$S_0 = \sum_i \sum_j W_{ij}$$

If we row-standardise, so that each row of W sums to 1 then

$S_0 = N$ and thus $$I = \frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\varepsilon}' \hat{\varepsilon}}$$

which is equal to the slope of the regression of $W\hat{\varepsilon}$ on $\hat{\varepsilon}$

Calculating Moran's I in practice

- Given *I*, we need to compare it with what we would expect under the null hypothesis of no residual autocorrelation

$$E(I) = tr(MW)/(N-K)$$

$$M = I - X(X'X)^{-1}X'$$

$$Var(I) = \frac{tr(MWMW') + tr(MWMW) + [tr(MW)]^2}{(N-K)(N-K+2)} - (E(I))^2$$

- These are the moments we would expect if the residuals were independent draws from a normal distribution

Calculating Moran's I in practice

- The test statistic is Z, which has the following distribution under the null hypothesis

$$Z = \frac{I - E(I)}{\sqrt{Var(I)}} \sim N(0,1)$$

- if $Z > 1.96$ or $Z < -1.96$ then we reject the null hypothesis of no residual spatial autocorrelation
  - infer that there is spatial autocorrelation in the regression residuals
  - BUT there is a 5% chance of a Type I error, false rejection of the null
- In the case of our house price data, *I* is 19.96 standard deviations above expectation
- a very clear indication that there is positive residual spatial autocorrelation

Calculating Moran’s I in practice

- Positive spatial autocorrelation is when ‘nearby’ residuals tend to have take similar values
  - Eg above average positive residuals may cluster together
- Negative spatial autocorrelation would be when ‘nearby’ residuals tend to be different
  - Positive residuals tend to be surrounded by negative ones and vice versa
- There are several alternatives to Moran’s I, and Moran’s I may also detect things other than spatially autocorrelated residuals
  - Moran’s I will also tend to detect heteroscedasticity, that is when the residuals have different variances rather than a common variance.
- However it is the most well known method of detecting spatial autocorrelation in regression residuals.

Modelling spatial dependence

- Say we have a significant Moran's I static, what next?
- We need to eliminate the spatial dependence
- one way to do this is to introduce an spatial autoregressive lag (spatial lag model)
- Consistent estimation via maximum likelihood OR via two stage least squares, OLS is not consistent because of the endogeneity of *Wy*

$$y = X\beta + \varepsilon \longrightarrow y = \rho Wy + X\beta + \varepsilon$$

X is a N x k matrix

β is a k x1 vector

ε is an N x 1 vector of errors

ρ is a scalar parameter

W is an Nx N matrix

Spatial lag model

- Here I list the values of these variables for the first 10 of the UALADs.

district	uaname	y	Wy
1.0	Luton	87464	168313
2.0	Mid_Bedfordshire	138856	151526
3.0	North_Bedfordshire	117530	137574
4.0	South_Bedfordshire	126650	157673
5.0	Bracknell_Forest	167633	200166
6.0	Reading	150094	186756
7.0	Slough	126361	222769
8.0	West_Berkshire	209543	170172
9.0	Windsor_and_Maidenhead	273033	183066
10.0	Wokingham	203059	205737

We can check whether Wy is a significant variable by adding it to our model

$$y = \rho W y + X\beta + \varepsilon$$

Dependent variable *y*	Spatial lag ML	
	estimate	t ratio
Constant *(X_1)*	-541.135534	-8.02
Local income *(X_2)*	393.33	5.58
Within-commuting-distance income *(X_3)*	27.45	6.89
Schooling quality *(X_4)*	149842.21	8.61
Number of households *(X_5)*	-0.35	-4.10
Spatial lag *(Wy)*	0.6089	14.90
R^2 adjusted	0.6330	
Standard Error	32.13	
Degrees of freedom	347	

Dependent variable *y*	ols	
	estimate	t ratio
Constant *(X_1)*	-571.874	-6.47
Local income *(X_2)*	864.0059	10.02
Within-commuting-distance income *(X_3)*	57.7055	14.08
Schooling quality *(X_4)*	175802.9235	7.74
Number of households *(X_5)*	-0.7112	-6.46
R^2 adjusted	0.567	
Standard Error	42.113	
Moran's I	0.39369	11.29
Degrees of freedom	348	

Created by demo_0.m

Fingleton B (2006) 'A cross-sectional analysis of residential property prices: the effects of income, commuting, schooling, the housing stock and spatial interaction in the English regions' *Papers in Regional Science* 85, 339-361

Direct, indirect and total effects in spatial lag model

- With *Wy*, the true effect of a variable, which typically is not the same as the regression coefficient β, as emphasized by LeSage and Pace (2009)
- the effects on dependent variable of a unit change in an exogenous variable , the derivative $\partial y/\partial X$,is not simply equal to the regression coefficient β
- the true derivative also takes account of the spatial interdependencies and simultaneous feedback embodied in the model, leading to a total effect that differs somewhat (typically) from β
- This derivative is somewhat complicated because it depends on the individual observations but can be represented by a mean
- See also Corrado and Fingleton(2012)

Corrado L. & Fingleton B. (2012) Where is the economics in spatial econometrics? Journal of Regional Science 52 210-239

Direct, indirect and total effects in spatial lag model

- Total effect = direct effect + indirect effect
- So we can partition the (average) total effect into a direct and an indirect effect
- The direct effect
  - gives the effect of *X* on *y* when the locations of *X* and *y* are the same
  - Differs from β because at location *r*, a change in *X* affects *y*, which then affects *y* at location *s* (*s n.e. r*) and so on, cascading through all areas and coming back to produce an additional effect on *y* at *r*
- Indirect effect = total effect – direct effect
  - The effect of *X* on *y* when *X* and *y* are not in the same location

Direct, indirect and total effects given lagged dependent variable
ML estimates : spatial lag model

Variable	Coefficient	Asymptot t-stat	z-probability
const	-541.135534	-8.023904	0.000000
local_income	393.326396	5.577120	0.000000
commuting_income	27.450614	6.894253	0.000000
supply	-0.353572	-4.104351	0.000041
schooling	149842.210059	8.613959	0.000000
rho	0.608979	14.898820	0.000000

Direct	Coefficient	t-stat	t-prob
local_income	439.507710	5.895381	0.000000
commuting_income	30.341573	7.244365	0.000000
supply	-0.394318	-4.179838	0.000037
schooling	167313.290102	8.807929	0.000000

Indirect	Coefficient	t-stat	t-prob
local_income	578.415728	5.010304	0.000001
commuting_income	39.749552	6.981685	0.000000
supply	-0.520811	-3.590056	0.000377
schooling	222208.985696	4.900776	0.000001

Total	Coefficient	t-stat	t-prob
local_income	1017.923438	5.870218	0.000000
commuting_income	70.091125	8.466974	0.000000
supply	-0.915129	-3.999705	0.000077
schooling	389522.275797	6.582510	0.000000

Created by demo_0.m

The spatial Durbin model: a 'catch all' spatial model

This includes a spatial lag *Wy* and a set of spatially lagged exogenous regressors WX

$$y = \rho Wy + X\beta + WX\gamma + \varepsilon$$

y = the dependent variable, an N x 1 vector
Wy = the spatial lag, an N x 1 vector
X = an N x K matrix of regressors, with the first column equal to the constant
β = a K x 1 vector of regression coefficients
ρ = the spatial lag coefficient
ε = an N x1 vector of errors
WX is the N by K matrix of exogenous lags resulting from the matrix product of W and X
γ is the corresponding coefficient vector.

Restricting the parameters of the spatial Durbin leads back to the spatial lag model or to the spatial error model spatial Durbin model : ML estimates

Created by demo_0.m

Variable	Coefficient	Asymptot t-stat	z-probability
const	-513.835677	-4.146915	0.000034
local_income	-7.730616	-0.083091	0.933780
commuting_income	40.795703	6.257112	0.000000
supply	-0.103221	-1.106877	0.268347
schooling	134249.627896	7.733356	0.000000
Wlocal_income	974.661531	6.096601	0.000000
Wcommuting_income	-25.325850	-3.358633	0.000783
Wsupply	-0.496569	-3.109303	0.001875
Wschooling	8596.323682	0.265708	0.790464
rho	0.621996	13.257551	0.000000

```
Rbar-squared      =   0.6549
Standard Error    =  873.1750^0.5 = 29.55
```

Special cases of the spatial Durbin

spatial lag

$$y = \lambda Wy + X\beta + WX\gamma + \varepsilon$$

if $\gamma = 0$

then $y = \lambda Wy + X\beta + \varepsilon$

spatial error

$$y = \lambda Wy + X\beta + WX\gamma + \varepsilon$$

if $\gamma = -\lambda\beta$

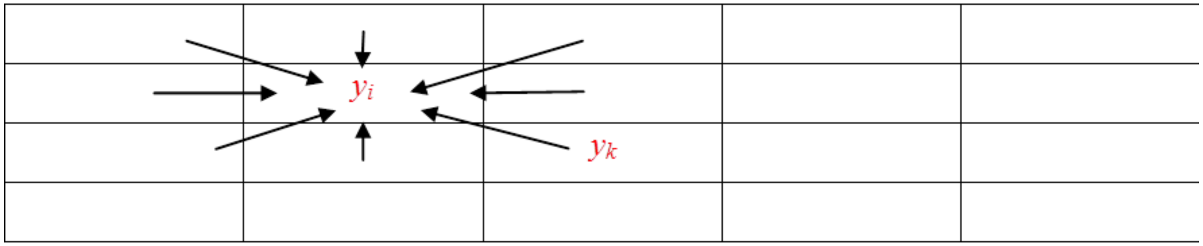
then $y = X\beta + \varepsilon$

and $\varepsilon = \lambda W\varepsilon + u$

Endogeneity of the spatial lag

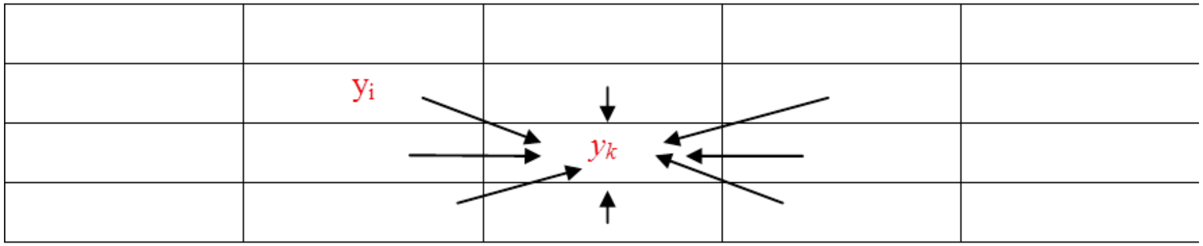
$$y = \lambda W y + X\beta + \varepsilon$$

y_i depends on Wy hence y_k



y_k (part of Wy) depends on y_i

so Wy depends on y_i and hence ε_i



Endogeneity of the spatial lag

$$y = \lambda W y + X\beta + \varepsilon$$

- y_i depends on all other y s, including y_k because they are within Wy.
- But y_k also depends on all other y s, including y_i because they are within Wy.
- So Wy determines y_i and is determined by it.

- So we have to use the appropriate likelihood function or 2sls to obtain consistent estimates.

$$y_i = f\left(\sum_{j=1}^{N} W_{ij} y_j\right)$$

$$y_k = f\left(\sum_{j=1}^{N} W_{kj} y_j\right)$$

- there are problems estimating these models by OLS
  - With the spatial lag model, the parameter estimates are biased
  - With the spatial error model, the parameter standard errors and hence the t-ratios are biased
- There are some appropriate (i.e consistent) estimators
- ML (maximum likelihood)
- 2sls/IV/GMM

Two stage least squares (2sls or TSLS)

- does not assume an explicit probability distribution for the errors so robust to non-normality
  - But not asymptotically the most efficient, ML more efficient when errors are normal, efficiency depends on instruments chosen
- avoids some of the computational problems of ML
- Allows several endogenous right hand side variables
- Consistent estimates, so plim of estimates are true values
- It is a familiar approach, being identical to 2sls in mainstream econometrics

Solving the problem

- Endogeneity lead to inconsistent OLS estimation
- Use an instrumental variables (IV) or equivalently two-stage least squares (2sls)
  - this involves replacing the endogenous variable(s) *X*, *Wy* (which are correlated with the error term) by 'proxy' variables. To do this we make use of (one or more) instrumental variable, that is independent of the error term.

Some conditions for a valid instrument

- Let *Wy* denote an endogenous variable (*X* could also be endogenous)
- Choose instrument (*Q*)
- Instrument relevance: corr(Q, Wy) $\neq 0$
- Instrument exogeneity: corr(Q, ε_i) $= 0$

- *Q* may be a single variable or a set of instruments hence a matrix

Inference using 2sls

- Statistical inference proceeds in the usual way.
- The justification is (as usual) based on large samples
- In large samples, the sampling distribution of the 2sls estimator is normal.
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way, e.g. estimated coefficient value ± 1.96SE
- This all assumes that the instruments are valid
- Note however that the standard errors from the second-stage OLS regression are not valid, because they do not take account of the fact that the first stage is also estimated
- So it is necessary to use a dedicated regression package that carries out 2sls with correct standard errors and hence t-ratios, rather than do two separate OLS regressions manually

software

- MATLAB
- Advantages
  - Lots of free spatial econometrics software available
      - E.g. James LeSage website
      - http://www.spatial-econometrics.com/
    - Ideal for innovative programming
    - Great graphics (eg maps via arc_histmap.m)
- Disadvantages
  - Complex

software

- MATLAB
- Availability
  - **Buy MATLAB and Simulink Student Version for £55**
    - Does this include add-ons?
    - Econometrics, Financial, Optimisation and Statistics toolboxes
    - http://www.mathworks.co.uk/programs/nrd/buy-matlab-student-version.html?ref=ggl&s_eid=ppc_3749
    - Training available but not specifically spatial econometrics
    - http://training.cam.ac.uk/ucs

software

- Stata
- Advantages
    - Familiar to many applied economists
        - » An economists package, not a general package for scientists
    - Easy to use interface
    - Youtube video ‘spatial econometrics in Stata’
    - http://www.youtube.com/watch?v=t7ADnMffink
  - Material becoming available
    - Maurizio Pisati Stata commands
      - http://www.stata.com/meeting/germany12/abstracts/desug12_pisati.pdf
    - Eg spatreg ado file
    - Mapping
      - http://www.stata.com/support/faqs/graphics/spmap-and-maps/

software

- Stata
- Disadvantages
  - Not so well developed at MATLAB
  - Packaged black-box approach allows standard methods but nothing novel (without real in depth Stata knowledge)

software

- Stata
- Availability

STATA/IC™

Perpetual licence PDF Documents on installation DVD	£120.00	GRADSIP
Annual licence PDF Documents on installation DVD	£63.00	GRADSIP
6 month licence PDF Documents on installation DVD	£45.00	GRADSIP

- Training available but not specifically spatial econometrics
- http://training.cam.ac.uk/ucs

software

- R
- Advantages
  - Free!
  - Many independent software writers
  - Becoming a favourite open source packages among economists, statisticians
- Disadvantages
  - A bit more complex than Stata but does more
  - Takes a while to get used to!

availability

- R
  - Youtube video ‘spatial econometrics in R’
  - http://www.youtube.com/watch?v=NLyjdmyokio
  - Material becoming available
    - Eg Install packages spdep

Course material on my webpages

- http://www.cantab.net/users/bf100/
- Go to *Teaching*
- Go down to *MPhil PGR07*
  - Excel demo .xlsx file
  - Excel demo notes
  - Lecture slides

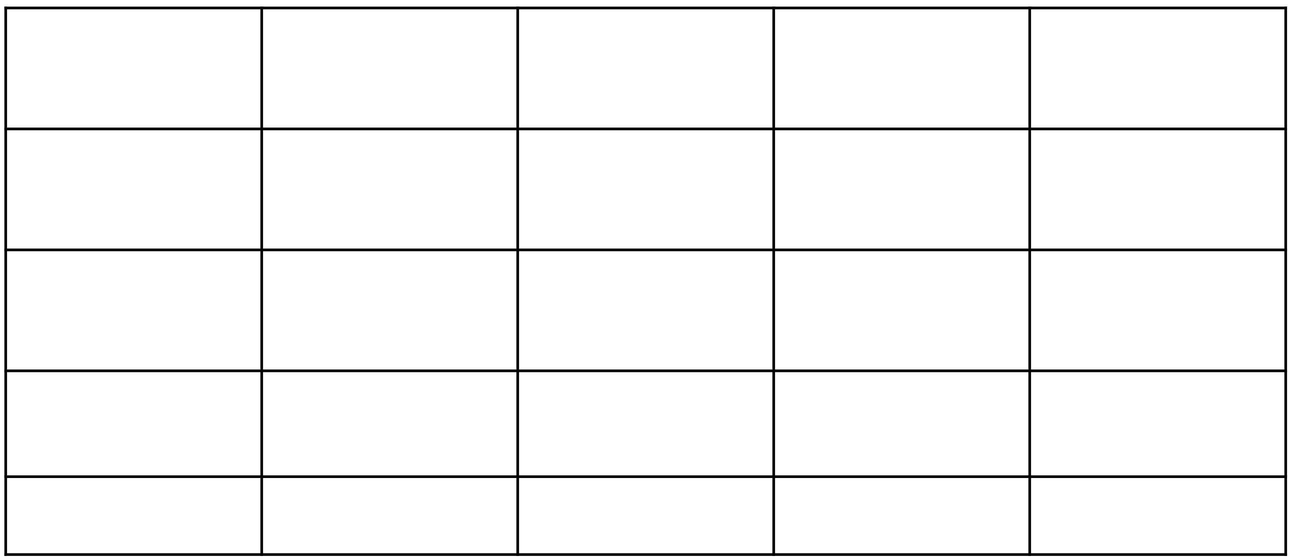
Running spatial regressions in Excel

Note of caution:

a) this is just to demonstrate how far one might get with Excel.
b) Some of the estimation is strictly inappropriate because it applies OLS which is an inconsistent estimator with an endogenous spatial lag. While consistent 2sls estimation can be carried out, this is left as an exercise for the student.
c) Likewise, the Moran's I analysis is informal and dedicated software should be used to carry out inference.
d) Excel is definitely not the preferred software for spatial econometrics. Here we are dealing with a simple problem involving 25 regions, and hence a 25 by 25 W matrix. Doing the same with more regions (eg 250) would be somewhat more difficult.

Running spatial regressions in Excel

25 square regions



Running spatial regressions in Excel

Ensure that Data Analysis can be seen on the extreme right of the Data tab, if NOT then File..options….add ins…manage Excel addins….tick Analysis Toolpack and Analysis Toolpack- VBA

1. **Open Excel_demo_c.xslm, sheet W (or Excel_demo.xslx if macros not available)**
This is a contiguity matrix for a 5 by 5 lattice (25 regions)

2. **Run Macro1** (in Excel tab at top, view, macros) **OR**
a) In sheet W, Select the cells a1 to 25y, replace selected cells in top left hand corner by the letter W, hit return
b) In sheet yx1x2, select the cells a1 to a25, replace selected cells in top left hand corner by the letter y, return
c) In sheet yx1x2, select the cells b1 to c25, rename as Xs
Now create the spatial lag Wy as the matrix product of W and y
d) Click on fx, select MMULT (found in math & trig)
For array 1 type W, for array 2 type y
Hold down shift (up arrow) +control(Ctrl) +return (left arrow) simultaneously
{} should appear around the command
Hit return and the matrix product of W and y will appear in column D

Running spatial regressions in Excel

3. **Run regression_1**

This regresses y on x1 and x2, putting the residuals in a column

4. **Run resWres**

This creates a vector of the spatial lags of the residuals (Wresids), so that we can then regress the lagged residuals (Wresids) on resids to find the value of Moran's I.
The method is the same as in 2d) this time using W and resids, thus creating the column Wresids.

5. **Run resreg1**

This is the regression of Wresids on resids, giving the Moran's I statistic equal to the slope. So in this case Moran's I = 0.6782. Note that we cannot strictly use the t ratio to test the significance of I

Running spatial regressions in Excel

6. **Run regression_2**

This regresses y on x1, x2 plus Wy, so we try to account for the spatially autocorrelated residuals by including the spatial lag Wy

Notice that the coefficient on the spatial lag Wy is equal to 0.7544, so it appears to be significant. However strictly we should be estimating this model by ML or 2sls because of the endogeneity of Wy. The residuals from this regression are created, which we call res2_.

7. **Run res2Wres2**

This forms a column of the spatial lag of the residuals (Wres2) so that the regression of Wres2_ on res2 can be carried out. Here we expect to see the extent of spatial autocorrelation is reduced because of the presence of Wy in the regression creating res2_.

8. Run resreg2

This is the regression of Wres2_ on res2. The slope gives a new measure of residual spatial dependence which can approximately be compared to Moran's I. In this case it is equal to the much smaller value of -0.1444.

Doing spatial econometrics in MATLAB

- Commands held in file demo_2sls.m
- Using same data, we first fit OLS regression
- This gives β coefficients
- for x1 $\beta_1 = 1.947$ (true value 2)
- for x2 $\beta_2 = 3.246$ (true value 3)
- Moran's I =0.33771
- These values are the same as obtained using Excel
- But E(I) and var(I) also calculated

Doing spatial econometrics in MATLAB

demo_2sls.m

Ordinary Least-squares Estimates
Dependent Variable =   y
R-squared   =   0.7490
Rbar-squared   =   0.7262
sigma^2   =   76.5926
Durbin-Watson =   1.9366
Nobs, Nvars   =   25,   3
****

Variable	Coefficient	t-statistic	t-probability
const	39.029834	8.246207	0.000000
x1	1.947332	6.061113	0.000004
x2	3.246412	5.657376	0.000011

morans i =0.33771
null morans i =-0.040533
morans i variance =0.023782
z statistic =2.4527
p-value =0.014178

Moran's *I* <u>same</u> as using Excel
But now we can carry out a valid
Test of its significance.
It is significantly greater than
expected under Null hypothesis of
no residual spatial autocorrelation

- OLS with spatial lag Wy
  - β coefficients
  - for x1 β_1 = 1.978650 (true value 2)
  - for x2 β_2 = 2.933007 (true value 3)
  - For Wy, ρ = 0.614303 (true value 0.5)
- Here because we are using OLS rather than a consistent estimator, the coefficient estimates are biased

Doing spatial econometrics in MATLAB

demo_2sls.m

OLS for spatial lag model, biased estimates

Ordinary Least-squares Estimates
Dependent Variable = y
R-squared = 0.8411
Rbar-squared = 0.8184
sigma^2 = 50.8027
Durbin-Watson = 1.9787
Nobs, Nvars = 25, 4
****

Variable	Coefficient	t-statistic	t-probability
Wy	0.614303	3.488306	0.002192
const	-6.223350	-0.459851	0.650349
x1	1.978650	7.557463	0.000000
x2	2.933007	6.163027	0.000004

Estimates same as obtained by Excel, suggesting significant spatial lag

Doing spatial econometrics in MATLAB

- 2sls for spatial lag model
  - β coefficients
  - for x1 β_1 = 1.974524 (true value 2)
  - for x2 β_2 = 2.974293 (true value 3)
  - For Wy, ρ = 0.533379 (true value 0.5)
- Unbiased estimates because 2sls is a consistent estimator
- The instruments are the spatial lag of the exogenous variables x1 and x2, and the spatial lag of the spatial lag
- Instruments
  - WX, WWX

Doing spatial econometrics in MATLAB

2sls estimates for spatial lag model

demo_2sls.m

Two Stage Least-squares Regression Estimates
Dependent Variable = y
R-squared = 0.8395
Rbar-squared = 0.8165
sigma^2 = 51.3135
Durbin-Watson = 2.0000
Nobs, Nvars = 25, 4
****

Variable	Coefficient	t-statistic	t-probability
Wy	0.533379	2.686647	0.013813
const	-0.261988	-0.017317	0.986348
x1	1.974524	7.502932	0.000000
x2	2.974293	6.190169	0.000004

Unbiased estimate

The end

- Thanks for your attention!