11th Economics Summer Seminars

Pamukkale University

Denizli

Turkey

Applied Spatial Econometrics

Bernard Fingleton

University of Cambridge

UK

- In previous lecture we have seen how spatially autocorrelated residuals obtained from a cross-section regression can lead to variants of the spatial Durbin model

spatial lag

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

if $\gamma = 0$

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

spatial error

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

if $\gamma = -\lambda\beta$

then $y = X\beta + \varepsilon$

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

- We showed that 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
- We now consider appropriate (i.e consistent) estimators
- ML (maximum likelihood)
- 2sls/IV/GMM

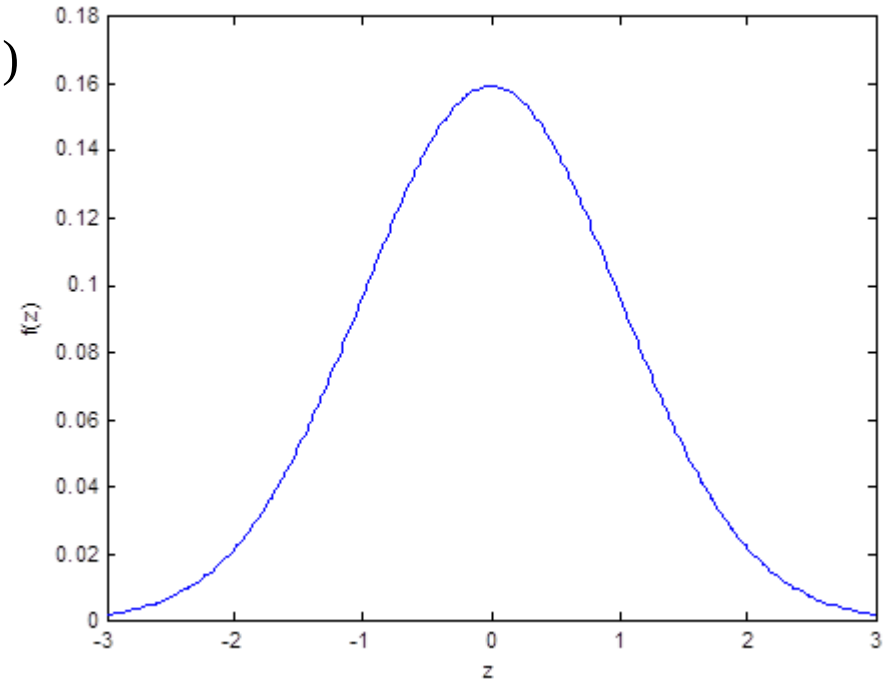
Maximum likelihood

probability density function for $x \sim \mathrm{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

$$z = \frac{x-\mu}{\sigma}$$

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-z^2}{2}\right\}$$



Consider now a small number of realizations of *z*, let us say *z* = [-3 -2.9 2.9 3]'. The likelihood of these occurring is small. In fact it is the product $L = f(z_1) f(z_2) f(z_3) f(z_4)$

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-z_i^{\,2}}{2}\right\} = \frac{1}{\sigma^n 2\pi^{n/2}} \exp\left\{\frac{-z'z}{2}\right\}$$

z_i	$-z_i^2$	$\exp(-z_i^2)$		
-3.0	-9	0.0001234		
-2.9	-8.41	0.0002226		
2.9	-8.41	0.0002226		
3.0	-9	0.0001234		
	$\Pi_i \exp(-z_i^2) =$	7.5486e-016	$-z'z =$ -34.8200	$\exp(-z'z) =$ 7.5486e-016

the realization z = [-0.2 -0.1 0.1 0.2]' is much more likely

z_i	$-z_i^2$	$\exp(-z_i^2)$		
-0.2000	-0.04	0.9608		
-0.1000	-0.01	0.9900		
0.1000	-0.01	0.9900		
0.2000	-0.04	0.9608		
	$\Pi_i \exp(-z_i^2) =$	0.9048	$-z'z =$ -0.1000	$\exp(-z'z) =$ 0.9048

Likelihood for OLS residuals ε

$$L = \frac{1}{\sigma^n 2\pi^{n/2}} \exp\left\{\frac{-\varepsilon'\varepsilon}{2\sigma^2}\right\}$$

σ^2 is the residual variance

One set of residuals will give one value L, another set another L value

Maximum likelihood estimation is that set of residuals (in other words regression coefficients leading to the residuals) that maximises *L*.

Assuming normal errors, this is precisely the same as minimising the sum of the squared residuals, ie OLS.

Spatial error model

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

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

$$u \sim N(0, \sigma^2 I)$$

so that

$$u = (I - \lambda W)\varepsilon = A\varepsilon$$

the joint likelihood function is

$$L = |A| \frac{1}{\sigma^n 2\pi^{n/2}} \exp\left\{ \frac{-(A\varepsilon)'(A\varepsilon)}{2\sigma^2} \right\}$$

Notice we now have $|A|$ known as the determinant of *A*, which is sometimes known as the Jacobean of the transformation from u to ε; In fact *A* is the matrix of partial derivatives $\frac{\partial u_j}{\partial \varepsilon_i}$ and ε is a set of residuals for a given set of parameters of the model $y = X\beta + \varepsilon$ (ie β, λ, σ), so we choose parameters β, λ, σ that maximise *L*.

Spatial lag model

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

$$\varepsilon \sim N(0, \sigma^2 I)$$

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

$$\varepsilon = (I - \rho W) y - X\beta$$

$$\varepsilon = Ay - X\beta$$

the joint likelihood function is

$$L = |A| \frac{1}{\sigma^n 2\pi^{n/2}} \exp\left\{ \frac{-(Ay - X\beta)'(Ay - X\beta)}{2\sigma^2} \right\}$$

Now A is the Jacobean of the transformation from ε to y. As above the maximum likelihood estimates are the set of values of β, ρ, σ^2 that give the maximum value of L.

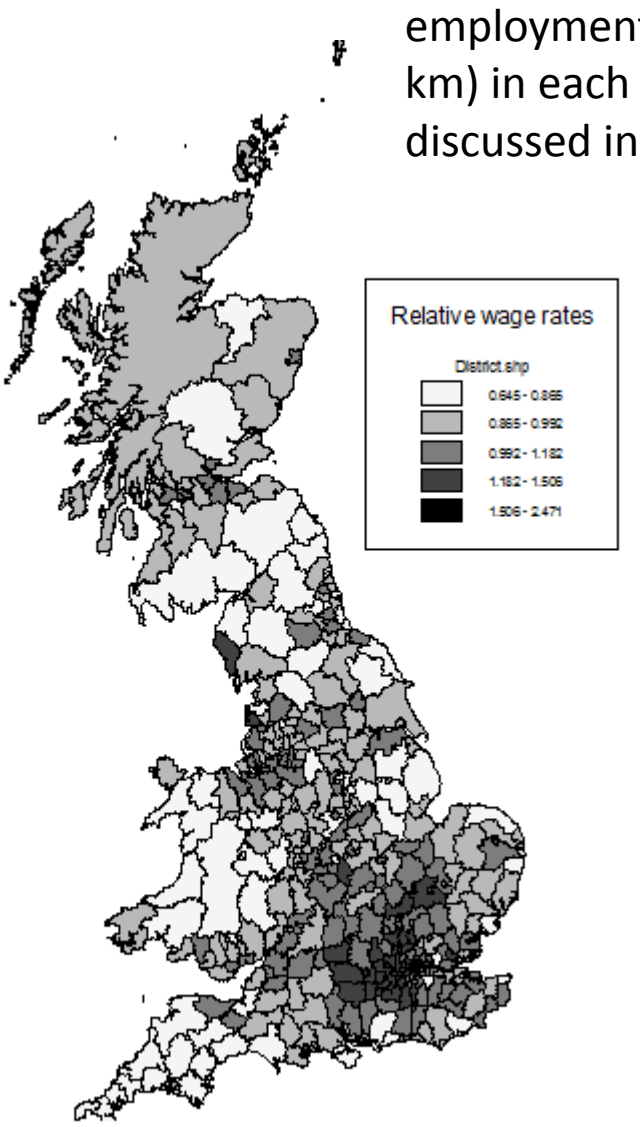
likelihood ratio tests

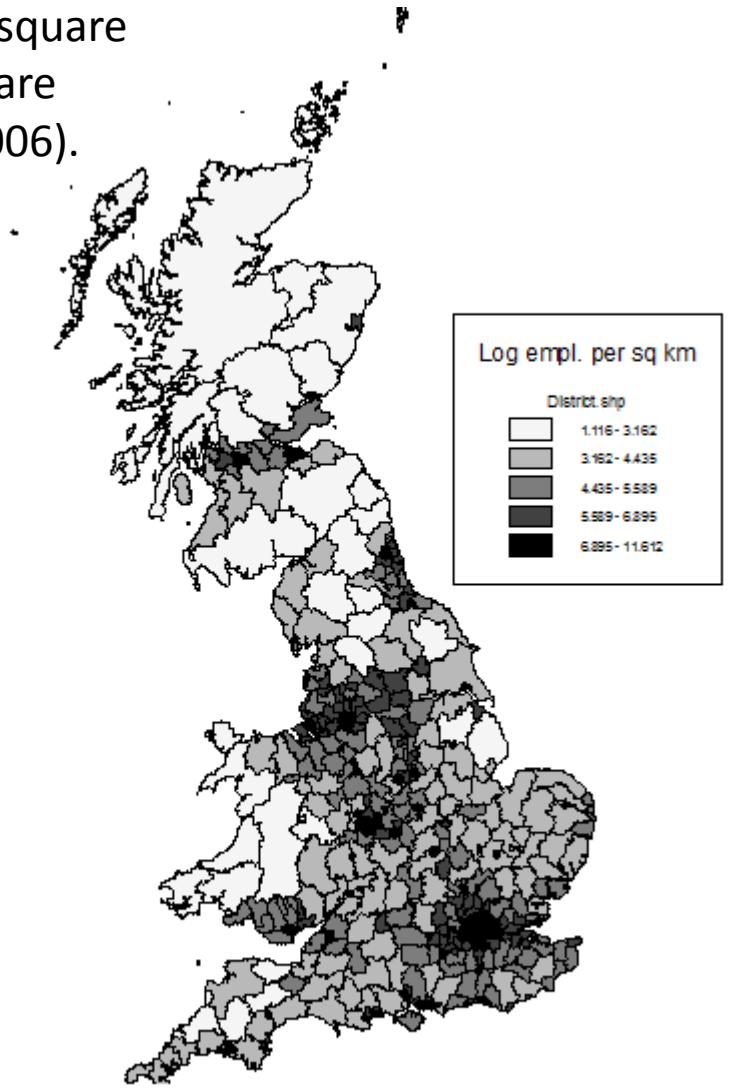
- The significance of either individual parameter estimates or of sets of parameter estimates can be assessed by likelihood ratio tests
- a pair of nested models is compared, the second of which is identical to the first apart from k >=1 restrictions that have been placed on the parameters of the first model
- The test essentially places on scale of zero to one the ratio of the likelihoods of the two models
- The numerator is the likelihood of the model restricted under the null hypothesis, and is therefore at most equal to the denominator which is the likelihood of the unrestricted model
- In practice one works with log L_R and log L_U which are the natural logs of the restricted and unrestricted models

$$-2\log\left(\frac{L_R}{L_U}\right) = 2\log\left(\frac{L_U}{L_R}\right) = 2\left\{\log(L_U) - \log(L_R)\right\} \sim \chi^2_k$$

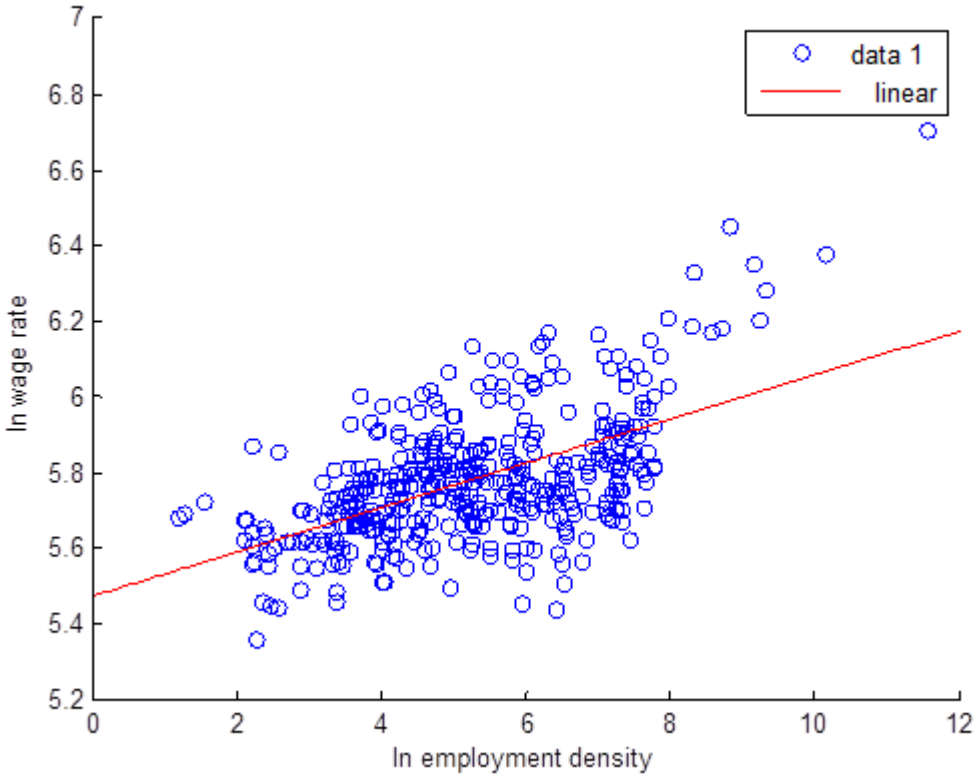
Under null hypothesis that the restrictions are true

model of wage rates by UALAD across GB. The dependent variable is the ln wage rate in each area, and this is dependent on the ln employment density (employment per square km) in each area. The data and theory are discussed in more detail in Fingleton(2006).





Fingleton B (2006) 'The new economic geography versus urban economics : an evaluation using local wage rates in Great Britain', *Oxford Economic Papers* 58 501-530


data 1
linear
ln wage rate
ln employment density
7
6.8
6.6
6.4
6.2
6
5.8
5.6
5.4
5.2
0
2
4
6
8
10
12

A) OLS model

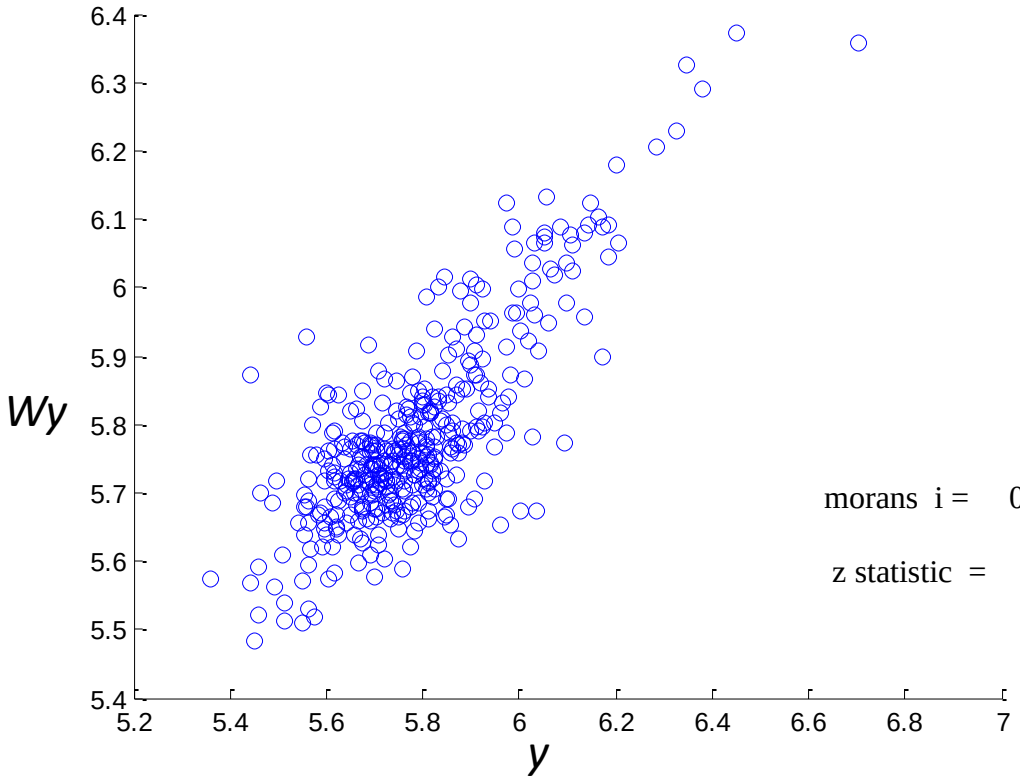
$$y = X\beta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

Created by demo_1.m

	$\hat{\beta}$	$s.e.$	$t = (\hat{\beta} - 0) / s.e.$	$p - value$
constant	5.4752363	0.023831146	229.75128	0
emp. density	0.058332136	0.004319197	13.50532	0

log-likelihood = 221.6148, df = n – k = 408 - 2 =406



morans i = 0.4150 null morans i = -0.0038 morans i variance = 0.0011

z statistic = 12.4824 p-value = 0

A portion of the matrix *W*

0	0	0	0.5	0
0	0	0.1667	0.1667	0
0	0.2	0	0	0
0.1428	0.1428	0	0	0

W has 408 rows and 408 columns, here we see the first 4 rows and 5 columns

These illustrate the hypothesised spatial interdependence of regions

Thus the value W(1,5) = 0.5 means that we assume there are two regions Interacting with region 1, since the matrix is standardised so that rows sum to 1

Likewise region4 interacts with 7 regions, since 7 * 0.1428 = 1.0

There are many alternatives to these hypothesised weights, but here only contiguous regions interact, otherwise the weight is 0

D) Unrestricted spatial Durbin model

Created by demo_1.m

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

$$\varepsilon \sim N(0, \sigma^2 I)$$

Variable	Coefficient	Asymptot t-stat	z-probability
const	1.885826	8.063784	0.000000
emp. density	0.019625	4.447737	0.000009
W*emp. density	0.014092	2.206810	0.027327
rho	0.643981	14.874570	0.000000

loglikelihood = 468.2947 df = N - k = 408 - 4 = 404

C) Spatial error model

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

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

$$u \sim N(0, \sigma^2 I)$$

Variable	Coefficient	Asymptot t-stat	z-probability
const	5.653089	167.098250	0.000000
emp. density	0.022036	5.003729	0.000001
lambda	0.748979	21.960853	0.000000

loglikelihood = 452.2114 df = N - k = 408 - 3 = 405

B) Spatial lag model

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

$$\varepsilon \sim N(0, \sigma^2 I)$$

Variable	Coefficient	Asymptot t-stat	z-probability
const	1.551050	7.848747	0.000000
emp. density	0.024102	7.012217	0.000000
rho	0.709987	20.026740	0.000000

loglikelihood = 465.5458 df = N - k = 408 - 3 = 405

Inference via Maximum likelihood

- commencing with our most complex model, the spatial Durbin model (D), we are more likely to have well behaved errors that are consistent with what is assumed for ML
  - This is referred to as the so-called top-down approach
- impose the restriction to reduce from D to C (the Spatial error model) gives

$$\text{restriction } \gamma = -\lambda\beta$$

$$2\left\{\log(L_U) - \log(L_R)\right\} = 2\{468.29 - 452.21\} = 32.17$$

$$32.17 > \chi^2_{1,0.01} = 6.63$$

The restriction leads to a significant loss of fit

Inference via Maximum likelihood

- reduce from D to B, reducing the unrestricted spatial Durbin model to the Spatial lag model

$$\text{restriction } \gamma = 0$$

$$2\left\{\log(L_U) - \log(L_R)\right\} = 2\{468.29 - 465.55\} = 5.5$$

$$5.5 < \chi^2_{1,0.01} = 6.63$$ The restriction can be maintained

- reduce from D to A, reducing the unrestricted spatial Durbin model to the OLS model

$$\text{restrictions } \gamma = 0, \rho = 0$$

$$2\left\{\log(L_U) - \log(L_R)\right\} = 2\{468.29 - 221.62\} = 493.36$$

$$493.36 > \chi^2_{2,0.01} = 9.21$$ The restrictions lead to a significant loss of fit

Some limitations of ML

- the calculation of the determinant of A is a major computational problem for data sets that are medium size, meaning 1000 areas and above
- Similarly there are problems in calculating the information matrix which provides the standard errors of the parameter estimates. This involves calculation of eigenvalues and inverses of large matrices that are not easily done for large samples
- the ML approach requires an explicit error probability distribution. A common assumption, which has been made here, is that the errors are normally distributed. However, this may not always be realistic
- The standard single equation methods only allow one endogenous variable, *Wy*, but we may wish to introduce additional endogenous variables

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

In general : 2 main reasons for endogeneity with cross-sectional data

1. Simultaneous equations bias
2. Omitted variables bias

3. Also we might have errors-in-variables
   - This is when we cannot measure the true *X* variable, so that there is uncertainty attached to the measured value
   - See Le Gallo J, Fingleton B (2012) 'Measurement errors in a spatial context' *Regional Science and Urban Economics* 42 114-125 for more on this….

1) Simultaneous equations bias

Endogenous spatial lag

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

$$corr(Wy, \varepsilon) \neq 0$$

generally

$$y_i = \beta_1 X_i + \varepsilon_i$$

$$X_i = \gamma_1 y_i + \nu_i$$

$$y_i = \frac{(\beta_1 \nu_i) + \varepsilon_i}{1 - (\beta_1 \gamma_1)}$$

$$X_i = \frac{(\gamma_1 \varepsilon_i) + \nu_i}{1 - (\beta_1 \gamma_1)}$$

2) Omitted variable bias

$$y_i = \beta_1 X_i + \beta_2 \tilde{W}_i + \varepsilon_i \qquad (True)$$

$$y_i = \beta_1 X_i + (\beta_2 \tilde{W}_i + \varepsilon_i)$$

$$y_i = \beta_1 X_i + v_i \qquad \text{(We estimate)}$$

If $Corr(X, \tilde{W}) \neq 0$ then $Cov(X, v) \neq 0$

3) Errors in variables

Suppose X_i is measured imprecisely by $\tilde{X}_i$ but we want to estimate the true relationship $y_i = b_0 + b_1 X_i + e_i$

In fact using $\tilde{X}_i$ the true relationship becomes

$$y_i = b_0 + b_1 \tilde{X}_i + [b_1 (X_i - \tilde{X}_i) + e_i]$$

since $b_1 \tilde{X}_i - b_1 \tilde{X}_i = 0$

Suppose we estimate

$$y_i = b_0 + b_1 \tilde{X}_i + \nu_i$$

The error term $\nu_i = b_1 (X_i - \tilde{X}_i) + e_i$ contains the difference $(X_i - \tilde{X}_i)$

If $\text{corr}(\tilde{X}_i, (X_i - \tilde{X}_i)) \neq 0$ then OLS estimator $\hat{b}_1$ from $y_i = b_0 + b_1 \tilde{X}_i + \nu_i$ is a biassed and inconsistent estimator of the true b_1 in $y_i = b_0 + b_1 X_i + e_i$

Solving the problem

- Endogeneity lead to inconsistent OLS estimation
- Ideally we should eliminate measurement error, introduce omitted variables, estimate a system of simultaneous equations etc.
- Often these solutions are not achievable in practice, thus……
- The solution is to use an alternative estimation method known as 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)
- Instrument relevance: $\text{corr}(Q, Wy) \neq 0$
- Instrument exogeneity: $\text{corr}(Q, \varepsilon_i) = 0$

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

Two Stage Least Squares (TSLS)

- Stage 1: Isolate the part of *Wy* that is uncorrelated with the error

We do this by regressing Wy on Q using OLS

$$Wy = \pi_0 + \pi_1 Q + v$$

because Q is uncorrelated with ε

$\pi_0 + \pi_1 Q$ is uncorrelated with ε

we don't know π_0 or π_1 but we have estimated them

so as to obtain the predicted values of Wy

$$W\hat{y} = \hat{\pi}_0 + \hat{\pi}_1 Q$$

Two Stage Least Squares (TSLS)

Stage 2: Replace *Wy* by the predicted values of *Wy*

Next regress y on $W\hat{y}$ (the predicted Wy from the first stage regression)

$$y = \rho W\hat{y} + X\beta + \varepsilon \qquad (2)$$

because $W\hat{y}$ is uncorrelated with ε in large samples

then ρ, β can be estimated consistently by OLS using

this second stage regression

In practice we do not need to carry out two stages, one stage is sufficient to get the same answer, making use of the projection matrix P_H

$$W\hat{y} = Q(Q'Q)^{-1}Q'Wy = P_H Wy$$

Q is an N x q matrix of instruments. Matrix Q includes at least one instrument for the endogenous variable Wy, together with the other exogenous variables in the matrix X, including the constant, which act as instruments for themselves.

The parameter estimates from the second stage regression are given by the 2sls estimator

two stages $\hat{Z} = [W\hat{y}, X]$ $\quad \hat{\gamma}_{2sls} = \left[\hat{Z}'\hat{Z}\right]^{-1}\hat{Z}'y = [\hat{\rho}, \hat{\beta}]$

single stage $Z = [Wy, X]$ $\quad \hat{\gamma}_{2sls} = [Z'P_H Z]^{-1}Z'P_H y = [\hat{\rho}, \hat{\beta}]$

estimated standard errors as the main diagonal of the asymptotic variance matrix given by

$$Var(\hat{\gamma}_{2sls}) = \hat{\sigma}^2[Z'Q(Q'Q)^{-1}Q'Z]^{-1}$$

$$\hat{\sigma}^2 = (y - \hat{y})'(y - \hat{y}) / (N - K - G)$$

$$\hat{y} = Z\hat{\gamma}_{2sls}$$

where K is number of exogenous variables (including the constant) and G is the number of endogenous explanatory variables (here $G = 1$, corresponding to Wy). Notice here that in order to obtain the fitted values $\hat{y}$ we apply the 2sls parameter estimates to Z and not to $\hat{Z}$

Instruments recommended by Kelijian and Prucha(1998)

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

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon$$

$$E[y] = (I - \rho W)^{-1} X\beta$$

They show that this 2SLS estimator is consistent under some general regularity conditions

$$(I - \rho W)^{-1} X\beta \approx X\beta + \rho WX\beta + \rho^2 W^2 X\beta + \ldots$$

recommended instruments $Q = [\, X,\ WX, W^2 X \,]$ not correlated with ε

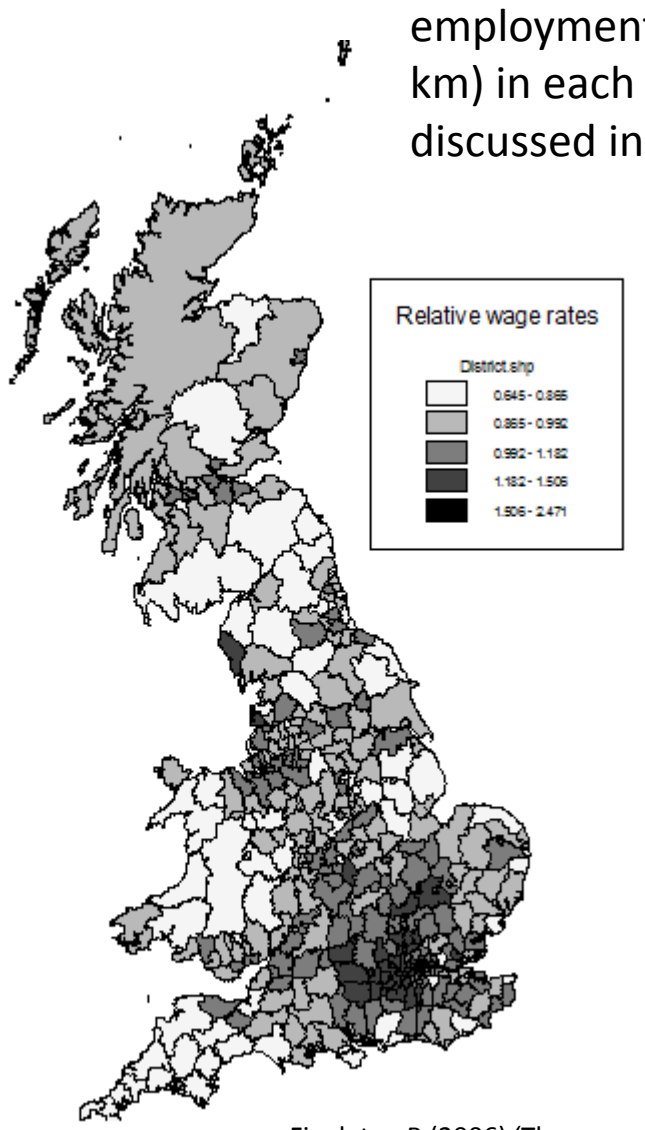
Lee (2003) argue that in cross-section spatial autoregressive model, the optimal instruments are

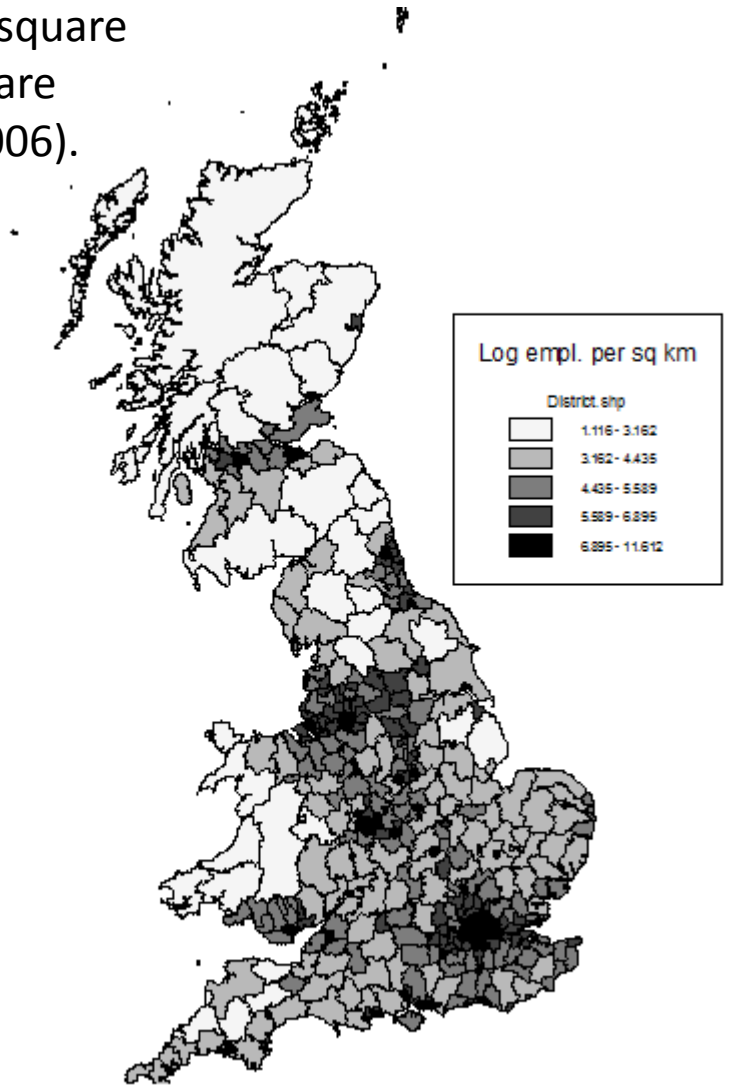
$$E(X, Wy) = (X, WE(y)) = (X, W(I - \rho W)^{-1} X\beta)$$

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/TSLS 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

model of wage rates by UALAD across GB. The dependent variable is the ln wage rate in each area, and this is dependent on the ln employment density (employment per square km) in each area. The data and theory are discussed in more detail in Fingleton(2006).





Fingleton B (2006) 'The new economic geography versus urban economics : an evaluation using local wage rates in Great Britain', *Oxford Economic Papers* 58 501-530

D) Unrestricted spatial Durbin model

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

$$\varepsilon \sim N(0, \sigma^2 I)$$

Created by demo_1.m

Variable	Coefficient	Asymptot t-stat	z-probability
const	1.885826	8.063784	0.000000
emp. density	0.019625	4.447737	0.000009
W*emp. density	0.014092	2.206810	0.027327
rho	0.643981	14.874570	0.000000

loglikelihood = 468.2947 df = N - k = 408 - 4 = 404

C) Spatial error model

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

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

$$u \sim N(0, \sigma^2 I)$$

Variable	Coefficient	Asymptot t-stat	z-probability
const	5.653089	167.098250	0.000000
emp. density	0.022036	5.003729	0.000001
lambda	0.748979	21.960853	0.000000

loglikelihood = 452.2114 df = N - k = 408 - 3 = 405

B) Spatial lag model

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

$$\varepsilon \sim N(0, \sigma^2 I)$$

Variable	Coefficient	Asymptot t-stat	z-probability
const	1.551050	7.848747	0.000000
emp. density	0.024102	7.012217	0.000000
rho	0.709987	20.026740	0.000000

loglikelihood = 465.5458 df = N - k = 408 - 3 = 405

2sls estimates of spatial lag model

Created by demo_1.m

```
Two Stage Least-squares Regression Estimates

Dependent Variable =     y

R-squared      =    0.6469

Rbar-squared   =    0.6451

sigma^2        =    0.0102

Durbin-Watson  =    2.0804

Nobs, Nvars    =    408,     3

***************************************************************

Variable          Coefficient      t-statistic    t-probability

Wy                  0.807218        11.407593         0.000000

const               1.013644         2.589266         0.009965

emp. density        0.019414         4.215630         0.000031
```

$\hat{\sigma}^2 = 0.0102$

$\hat{\gamma}_{2sls} = \left[\hat{\rho}\ \hat{\beta}_1\ \hat{\beta}_2\right] = \left[0.807218\ 1.013644\ 0.019414\right]$ instruments WX, W^2X

$$\text{var}(\hat{\gamma}_{2sls}) = \begin{bmatrix} 0.0050 & -0.0277 & -0.0002 \\ -0.0277 & 0.1533 & 0.0013 \\ -0.0002 & 0.0013 & 0.0000 \end{bmatrix}$$

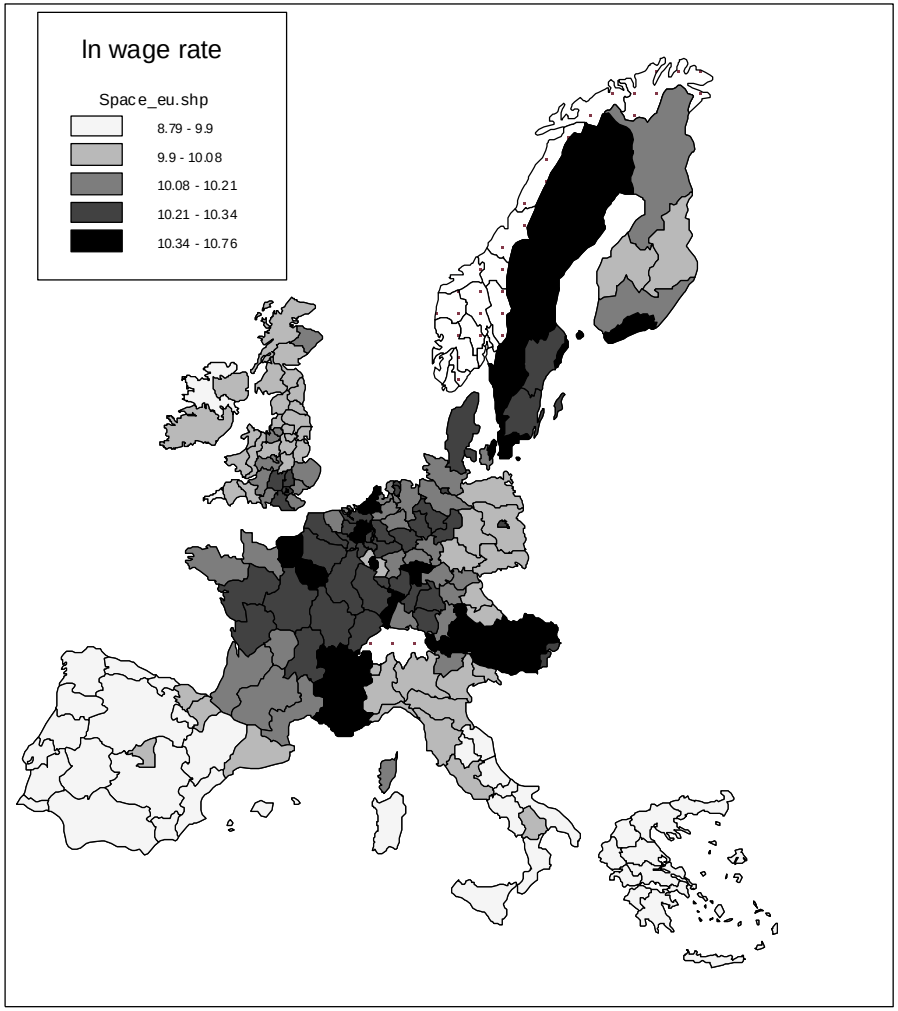
we have two goodness-of-fit statistics, R-squared and Rbar-squared. With OLS, R-squared is exactly equal to the square of the correlation between the dependent variable y and the fitted values $\hat{y}$. With 2sls this is not exactly true, but the correlation will give an approximate value. In both OLS and 2sls

$$R-squared = R^2 = 1 - \frac{Var(\varepsilon)}{Var(y)} = 1 - \frac{\hat{\sigma}^2}{Var(y)} = \frac{Var(y) - \hat{\sigma}^2}{Var(y)}$$

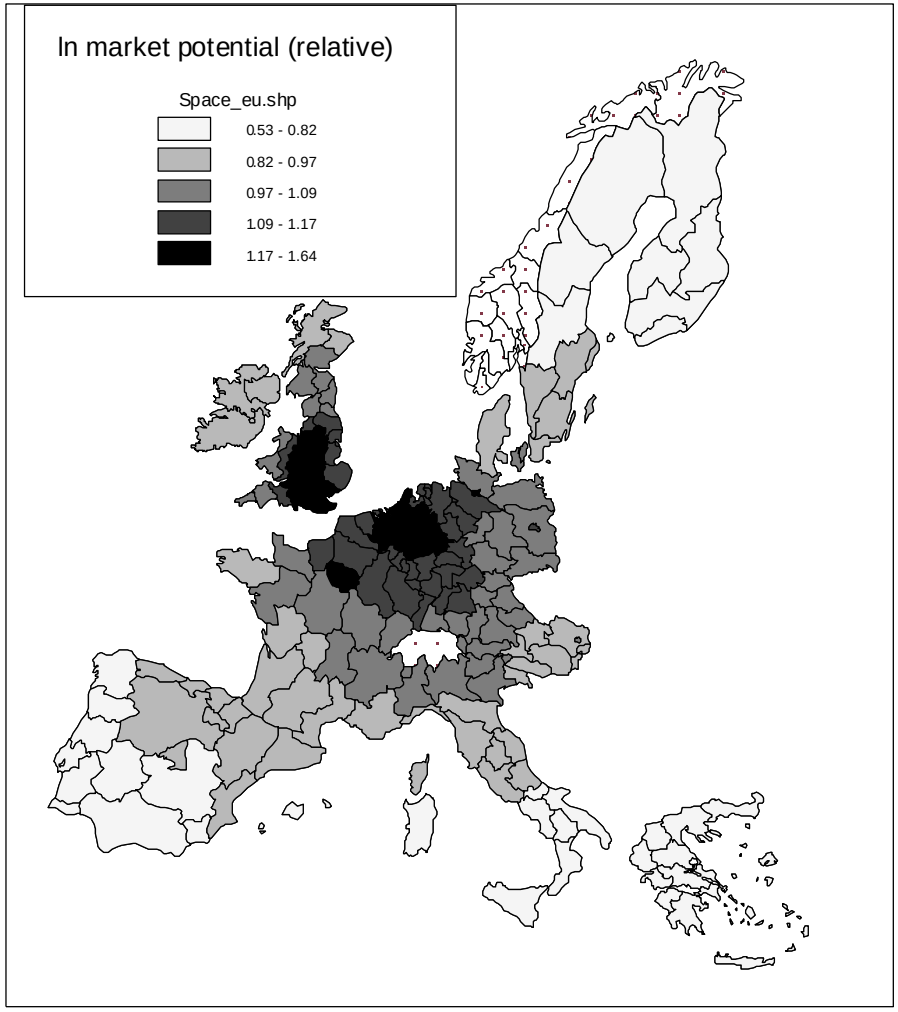
$$Rbar-squared = \bar{R}^2 = 1 - \frac{N-1}{N-K-G}(1-R^2)$$

R-squared is 'explained' variance of y, $Var(y) - \hat{\sigma}^2$ as a proportion of total variance of y $Var(y)$, and therefore is on a scale from 0 to 1. Because R-squared does not take account of how many variables there are, and may be high simply because there are lots of explanatory variables, it makes sense to control for this

Multiple endogenous variables



Data for 2003, taken from
Fingleton B, Fischer M (2010) Neoclassical Theory versus New Economic Geography. Competing explanations of cross-regional variation in economic development, Annals of Regional Science, 44 467-491


ln market potential (relative)
Space_eu.shp
0.53 - 0.82
0.82 - 0.97
0.97 - 1.09
1.09 - 1.17
1.17 - 1.64

What is market potential?

- Intuitively, it is the access to supply and demand at a particular location *i*.
- It depends on the on the level of income and prices in each area *i,j,k,l,m*….
- However remoter areas (eg *m*) add less to the market potential of location *i* because of transport costs between *m* and *i*.
- Where market potential is high, workers can bid up wage rates reflecting the advantages to producers in high market potential locations

Dependent variable *Y* = log(GVApw)

```
Model 2: OLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -2.51682      1.19136     -2.113    0.0356    **
  lnMP         1.28870      0.117013    11.01     2.66E-023 ***
```

Why is *ln MP* endogenous?

NEG (new economic geography) theory gives a

set on nonlinear simultaneous equations

involving wage rates w_i^M and market potential *MP*

wage rates depend on *MP*

but *MP* is partially determined by wage rates

New Entrants

Is there an omitted variable, also affecting wages?
We propose that 'New Entrants' should also be entered into our model

It simply takes the value 1 or zero according to whether a region is in a new entry country

Two reasons why ln*MP* and ε might be correlated

- Simultaneous equations bias
  - Market potential (ln*MP*) depends on wages (hence ε) hence corr(ln*MP,* ε) ≠0
- BUT also
- Omitted variables bias : *New Entrants*
  - *New Entrants* have low ln*MP,* so
  - corr(*New Entrants* ,ln*MP*) < 0
  - Since *New Entrants is* in ε,

    corr(ln*MP* ,ε) ≠0
- AND
  - *New Entrants possibly* depends on wages hence

    corr(*New Entrants* , ε) ≠0
- So to avoid ov bias, we need to introduce an extra 'endogenous' var *New Entrants*

Some instruments

- Q_1 =ln area of region in sq. km = ln(sqkm)
  - Sqkm is fixed, it is the area of the region and will not change in response to wage rates, or as a result of taking logs
  - Regions with smaller areas are cities, which are concentrations of economic activity with high market potential

- Q_2 =weighted average of log of areas of surrounding regions in sq. km = Wa(ln(sqkm))
  - Likewise, we do not alter the exogeneity by taking the weighted mean of ln(sqkm)
  - Having ‘cities’ nearby will add to an areas market potential

Some instruments

- Q_3 = log employment density
  - Equal to workers per square km
  - It may not be exogenous, we need to test this and the other instruments

Typically

- we have more than one rhs endogenous variable
- we want to use more than one instrumental variable

Why include additional instruments?

- with more than one instrument each the coefficient is said to be overidentified in this case
- In the case of just one instrument per endogenous variable, 2sls will work, we have in this case exact identification.
- but if we had less than one instrument per endogenous variable, then this would not work, the coefficients to be estimated would be underidentified

2sls with > 1 endogenous variable

- Assume that whether or not a country is a new entrant depends on its GVA per worker
- Then we have 2 endogenous variables. lnMP, new_entrant
- The 2 stages are as before but
- Take care that there are enough *instrumental* variables so as to avoid under-identification.
- With 3 instruments for our 2 endogenous variables we have overidentification
- can test for the validity of the instruments via the Sargan test

Created by demo_2.m

```
Two Stage Least-squares Regression Estimates
Dependent Variable =      lnGVApw
R-squared      =    0.8226
Rbar-squared   =    0.8212
sigma^2        =    0.0524
Durbin-Watson  =    1.0356
Nobs, Nvars    =    255,     3
***************************************************************
Variable         Coefficient      t-statistic    t-probability
lnMP               0.298355         2.840966         0.004865
new_entrant       -1.210214        -3.694551         0.000270
const              7.748651         7.001371         0.000000

instruments
ln_sqkm
WA_ln_sqkm
ln_empdens

endogenous vars
lnMP
new_entrant
```

Sargan test

- The Sargan test is a test of the validity of instrumental variables.
- It is a test of the overidentifying restrictions. The hypothesis being tested is
  - the instrumental variables are uncorrelated with the residuals
  - And therefore they are acceptable instruments.
- If the null hypothesis is confirmed statistically (that is, not rejected), the instruments pass the test; they are valid by this criterion.

Sargan test

- Instruments should be independent of the errors
- To test whether this is the case, we take the 2sls residuals as the dependent variable
  - 2sls residuals use the 2sls coefficient estimates and the original variables, not the instruments
- Then take the instruments (*Q*) and the other exogenous variables as regressors
- For valid instruments, the *Q* should be unrelated to the 2sls residuals
  - This assumes that the set of regressors is correct and there is no model misspecification
  - For instance we assume that the instruments do not have a direct effect, are not regressors

Sargan test : overidentifying restrictions

- Overidentification is when we have more Instruments than endogenous variables
- On its own each instrument will give a different estimate
- But we expect valid individual instruments to give more or less the same estimates
- If they differ, that suggests ‘something is wrong with the instruments’

Sargan test : overidentifying restrictions

- They are called ‘over-identifying restrictions’

because we test the null hypothesis that,

in the regression of the 2sls residuals depending on the regressors and Q,

the coefficients on the set of instruments (Q) can be restricted to zero

- This is what we would expect of all the instruments were valid, that is valid Q should be unrelated to the residuals

Sargan test : overidentifying restrictions

- It only works with over-identification, the test cannot be carried out with exact identification
  - If you have exact identification, and regress the instrument(s) on the 2sls residuals, the coefficient(s) is(are) exactly zero.
  - The same thing happens if you regress an exogenous variable on OLS residuals. By definition, the residuals are independent of the regressor, so you cannot test whether this is the case
- Thus we need more instruments than endogenous variables

Created by demo_2.m

Sargan test : overidentifying restrictions

Sargan test

LM test statistic is *N*R*^{2}
***N* is the number of regions**
***R*^{2} is from the regression of the X variables and instruments on the 2sls residuals**

LM test statistic = 0.318418

test statistic referred to χ^2
with degrees of freedom = 3-2 = 1

degree of overidentification = no. of instruments-no. of end. vars = 3 - 2

Hence 0.318418 has an upper tail probability = 0.57256 in χ^2_1

if tail probability below 0.05 ($\chi^2_{1,0.05} = 3.84$ **) then we reject the null and conclude that the instruments are endogenous and/or the model is misspecified**

Wu-Hausman : checking the exogeneity of variables

- An exogenous variable does not need to be instrumented, an endogenous one does
- Sometimes theory tells us that a variable is endogenous (eg *lnMP*)
- But we can also use diagnostics to tell us whether a variable is endogenous

Wu-Hausman : checking the exogeneity of variables

- The test, often referred to as the Wu-Hausman test, comprises 2 regressions
- Regression 1
  - (each possible) 'endogenous' variable is the dependent variable
  - the exogenous variables and the instruments Q are the independent variables,
  - save the *fitted values OR* the *residuals (*both give identical conclusions*)*

Wu-Hausman : checking the exogeneity of variables

- Regression 2
  - *y* variable is the dependent variable and 'endogenous', exogenous and *fitted endogenous variable values* (or *residuals*) are independent variables
  - If the *fitted endogenous variables* are significant, then they carry explanatory information additional to that that already contained in 'endogenous' and exogenous regressors
  - The *fitted endogenous variables* are exogenous and consistent by definition
  - If the 'endogenous' variables are also exogenous & consistent, then the *fitted endogenous variables* will be redundant
  - *If the fitted endogenous variables are* significant, this suggests that the 'endogenous' variables are not exogenous and consistent, and thus are truly endogenous

Created by demo_2.m

regressions for Wu-Hausman test

```
Model 9: OLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw

               coefficient   std. error   t-ratio     p-value
  ------------------------------------------------------------
  const          7.74865     1.09497        7.077    1.48E-011 ***
  new_entrant   -1.13052     0.0610374    -18.52     8.43E-049 ***
  lnMP           0.742390    0.173650       4.275    2.72E-05  ***
  fvMP          -0.444035    0.202362      -2.194    0.0291    **
  fv_ne         -0.0796919   0.329784      -0.2416   0.8093


Model 10: OLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw

               coefficient   std. error   t-ratio    p-value
  -----------------------------------------------------------
  const          7.12069     0.708099      10.06    3.20E-020 ***
  new_entrant   -1.22644     0.0458748    -26.73    1.56E-075 ***
  lnMP           0.360292    0.0692045      5.206   4.00E-07  ***

Comparison of Model 9 and Model 10:

  Null hypothesis: the regression parameters are zero for the variables

    fvMP
    fv_ne

  Test statistic: F(2, 250) = 2.87768, with p-value = 0.0581311
```

Wu-Hausman : checking the exogeneity of variables

- This indicates (p = 0.058) that the two variables *lnMP* and *new entrants* possibly are endogenous
- The outcome depends primarily on the significance of *fv_lnMP*
- It is likely that *new entrants* could be treated as exogenous

Weak instruments

- we wish to avoid weak instruments, which itself leads to bias and size distortion (Stock, Wright and Yogo, 2002, Stock and Yogo, 2005)
- our instruments should be sufficiently correlated with the endogenous regressors and while remaining orthogonal to the disturbances

Weak instruments

- Weak instruments can lead to serious problems in IV regression: biased estimates and/or incorrect size of hypothesis tests, with rejection rates well in excess of the nominal significance level

Weak instruments

- The first stage regressions indicate that while the instruments are strong for *lnMP* they are weak for *new entrants*
- *lnMP*
  - R-squared = 0.7281 Created by demo_2.m
  - Rbar-squared = 0.7248
- *new entrants*
  - R-squared = 0.0329
  - Rbar-squared = 0.0213
- However we treat *new entrants* as exogenous as a result of the Wu-Hausman test