Spatial Regression

3. Review - OLS and 2SLS

Luc Anselin

http://spatial.uchicago.edu
OLS estimation (recap)

non-spatial regression diagnostics

dendogeneity - IV and 2SLS
OLS Estimation (recap)
• **Linear Regression - Notation**

linear relationship between a dependent variable $y_i$ (at location $i$) and a set of explanatory variables $x_{ih}$, for $h = 1, ..., k$ subject to random error

- $y_i = \sum_h x_{ih} \beta_h + e_i$

  $e_i$ is random error term, with $E[e_i] = 0$, i.e., no systematic error
• **Linear Regression - Notation** (continued)

  in matrix notation, a $n$ times 1 column vector $y$ and a $n$ times $k$ vector $X$, with a $k$ times 1 coefficient vector $\beta$ and a $n$ times 1 random error vector $e$

  - $y = X\beta + e$

  - $E[e] = 0$
Conditional Expectation

under a set of regularity conditions (to follow)
the conditional expectation of $y$ given $X$ is linear in $X$

$$E[ y | X ] = E[ X\beta | X ] + E[ e | X ] = X\beta + 0$$

in other words, what would $y$ be on average if we knew $X$
• Marginal Effect

the effect of a change in $X$ on $y$

in a linear regression the marginal effect equals the regression coefficient (this is not the case in a nonlinear regression)

• $E[ y | \Delta X ] = \Delta X \beta$
• Selected Regularity Conditions

X non-stochastic (or if stochastic, with bounds on second moment) - the only randomness follows from the dependent variable y, any randomness in X is inconsequential

error term independent identically distributed (i.i.d), i.e., \( \text{Var } [e_i] = \sigma^2 \) or \( \text{E}[ee'] = \sigma^2 I = \) spherical error term

\( x_i \) and \( e_i \) uncorrelated for all i, i.e., signal (X) and noise (e) are not related
Ordinary Least Squares (OLS) Regression

under set of regularity conditions, yields the best (smallest variance) unbiased estimator = Gauss-Markov theorem

\[ b = (X'X)^{-1} X'y \]

\[ E[ b ] = E [ (X'X)^{-1} X'(X\beta) ] + E [ (X'X)^{-1} X'e ] = \beta \]

(since \( E[ X'e] = 0 \))
• OLS - Inference

with non-spherical errors

\[ \text{var}(b) = s^2 (X'X)^{-1} \]

\[ s^2 = \frac{e'e}{n - k} \]

an unbiased estimator of error variance

use in t-tests (with assumption of normality)
• Predicted Value

value of $y_i$ given $x_i$ using the estimates $b$

$$y_{ip} = \sum_h x_{ih} b_h$$

• Residual

difference between observed and predicted

$$u_i = y_i - y_{ip}$$

for regression with constant term $\text{avg}[u_i] = 0$

residual is NOT the same as the error term ($e_i$)
• General Covariance Structure

\[ y = X\beta + e, \quad E[ee^\prime] = \Omega \]

both heteroskedasticity and autocorrelation

\[ \text{Var}[\beta_{OLS}] = (1/n)(1/n X'X)^{-1}(1/n X'\Omega X)(1/n X'X)^{-1} \]

develop estimator for \((1/n)X'\Omega X\) (k by k) but NOT an estimator for \(\Omega\) (n by n)

example: White (sandwich) standard errors
Non-spatial Regression Diagnostics
Visual Diagnostics
• Predicted Value Map

shows spatial distribution of model prediction

a form of smoothing, i.e., what the model suggests $y$ should be, given the $X$ at each location
Predicted Value map - 1990 county homicide rates
(standard deviational map)
Residual Map

- High values (red) = model under-predicts \( (y > y_p) \)
- Low values (blue) = model over-predicts \( (y < y_p) \)
- Note extremes = poor fit of model
- Note spatial patterns, but visual inspection can be misleading
- Need for formal diagnostics
Residual map - 1990 county homicide rates (standard deviational map)
• Diagnostic Plots

use scatter plot function

plot residuals (y-axis) vs. predicted values (x-axis) as a visual diagnostic for heteroskedasticity

pattern should be more or less within the same range

“fan” or “flares” suggest heteroskedasticity, i.e., non-constant error variance
Residual Plot - Heteroskedasticity
• Caution

visual inspection of plots and maps can be misleading

use plots and maps to suggest additional variables for model

no substitute for formal specification diagnostics
Specification Tests
Regression Report - Diagnostics

<table>
<thead>
<tr>
<th>Test</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jarque-Bera</td>
<td>2</td>
<td>2833.409</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breusch-Pagan test</td>
<td>5</td>
<td>515.0796</td>
<td>0.0000000</td>
</tr>
<tr>
<td>Koenker-Bassett test</td>
<td>5</td>
<td>124.2749</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>20</td>
<td>242.8053</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

END OF REPORT
• Multicollinearity

• condition number

based on eigenvalues of $X'X$

“classic” rule of thumb = values $> 30$ suggest a problem

in practice, not taken too literally

in example, slightly over 30
• Normality

crucial assumption for exact inference

in practice, less crucial, since asymptotics (large sample theory) yield similar properties

Bera-Jarque(1981) test is based on third moment (skewness - asymmetry) and fourth moment (kurtosis - thick tails) of residuals

distributed as $\chi^2$ with two degrees of freedom

in example: 2833, thus highly non-normal
• Random Coefficients

a special form of heteroskedasticity

measure of fit of regression of squared residuals on squared explanatory variables


in example: both reject null very strongly
Heteroskedasticity

White (1980) test against heteroskedasticity of unknown form

measure of fit of regression of squared residuals on a polynomial in the explanatory variables

robust to many forms of misspecification

in example: rejects the null soundly
• Caveats

in large data sets (as in the south example), lack of normality is not a real problem

heteroskedasticity is hard to distinguish from spatial autocorrelation (and vice versa)

many types of misspecification (missing variables, nonlinearity) may lead to the rejection of a given null hypothesis, even when unrelated to that null hypothesis
Endogeneity
IV and 2SLS
Simultaneous Equation Bias
• Regularity Conditions

\[ E[x_i | e_i] = 0 \]
implies \( E[e] = 0 \) and \( E[X'e] = 0 \)
establishes unbiasedness

weaker \( \text{Cov}[X'e] \leq 0 \) for consistency

\[ \text{plim} \ (1/n) \ X'e = 0 \]
Simultaneous Equation Bias

\[ E[X'e] \neq 0 \]

OLS biased

\[ E[b] = E\{ (X'X)^{-1}X'(X\beta + e) \} = \beta + (X'X)^{-1}E(X'e) \neq \beta \]
• Simultaneous Equation Bias (2)

\[ \text{plim } [(1/n)X'e] \neq 0 \]

OLS inconsistent

\[ \text{plim}[b] = \beta + \text{plim}[(X'X/n)^{-1}] \cdot \text{plim}[X'e/n] \neq \beta \]
Instruments
General Model

\[ y = Z\theta + e \]

some \( z_{ki} \) correlated with \( e_i \)

two sources of variation: error and \( Z \)
• **Instruments**

new variable \( q \)

uncorrelated with errors \( E[ e_i \mid q_i ] = 0 \)

correlated with original \( Z \)
• **Regularity Conditions**

  instrument matrix $Q$, n by K, i.e., same dimension as $Z$

  instruments not multicollinear

  $\text{plim} \ (1/n) \ Q'Q = H_{QQ}$

  a finite, positive definite matrix
• Regularity Conditions (2)

  instruments correlated with $Z$

  $\text{plim } (1/n) Q'Z = H_{QZ}$, finite, rank $K$

  instruments uncorrelated with $e$

  $\text{plim } (1/n) Q'e = 0$, for each column of $Q$
IV Estimator
• Estimator

\[ \theta_{IV} = (Q'Z)^{-1}Q'y \]

consistency

\[ \theta_{IV} = (Q'Z)^{-1}Q'(Z\theta + e) \]

\[ \theta_{IV} = \theta + (Q'Z)^{-1}Q'e \]

\[ \text{plim}(\theta_{IV}) = \theta + \text{plim}(Q'Z/n)^{-1}\text{plim}(Q'e/n) \]

\[ = \theta + H_{QZ}^{-1}.0 = \theta \]
• Limiting Distribution

\[ \sqrt{n} (\theta_{IV} - \theta) = (Q'Z/n)^{-1}(Q'e/\sqrt{n}) \]

limiting distribution depends on \( Q'e/\sqrt{n} \)

requires central limit theorem to establish asymptotic normality

requires law of large numbers to establish consistency of asymptotic variance
• Limiting Distribution (2)

central limit theorem \( Q'e/\sqrt{n} \to N(0,M) \)

\[ M = \text{plim}[ (1/n) Q'ee'Q ] \]

\( ee' \) pertains to error variance covariance \( \Sigma \)

requires law of large numbers to yield

\[ M = \text{plim}[ (1/n) Q' \Sigma Q ] \]
• **Asymptotic Variance**

\[
\text{variance of } \sqrt{n} (\theta_{IV} - \theta) \text{ is variance of } (Q'Z/n)^{-1}(Q'e/\sqrt{n}), \text{ a constant (matrix) times a random variable (vector)}
\]

\[
\text{plim}(Q'Z/n)^{-1}. \text{plim}(Q' \Sigma Q/n].\text{plim}(Z'Q/n)^{-1}
\]

\[
= H_0QZ^{-1}MHZQ^{-1}
\]
• Asymptotic Distribution

$$\sqrt{n} \ (\theta_{IV} - \theta) \rightarrow N(0, H_{QZ}^{-1} M H_{ZQ}^{-1})$$

$$\theta_{IV} \rightarrow N(\theta, H_{QZ}^{-1} M H_{ZQ}^{-1}/n)$$

$$\text{Var}[\theta_{IV}] = \frac{1}{n}(Q'Z/n)^{-1}(Q'\Sigma Q/n)(Z'Q/n)^{-1}$$

Note: all n terms cancel out, s.t., variance can be estimated as

$$\text{Var}[\theta_{IV}] = (Q'Z)^{-1}(Q'\Sigma Q)(Z'Q)^{-1}$$
2SLS Estimator
• More Instruments than Endogenous Variables

instrument matrix $Q$ has rank $L > K$

$Q'Z$ no longer feasible - column ranks don’t match

$\text{plim}(Q'e/n) = 0$ holds for any linear combination of columns of $Q$

which instruments to choose?
Projection Matrix

could pick any K columns from Q

optimal choice is combination of columns of Q closest to Z

\[ Z_p = Q(Q'Q)^{-1}Q'Z \]

predicted values of least squares regression (projection) of Q onto columns of Z yields K “instruments”

\[ P_Q = Q(Q'Q)^{-1}Q' \]

is the projection matrix
2SLS Estimation

use $Z_p$ as matrix of instruments

$$\theta_{2SLS} = (Z_p'Z)^{-1}Z_p'y$$

$Z_p'Z = Z_p'Z_p$ same as OLS of $y$ on $Z_p$

$$\theta_{2SLS} = (Z_p'Z_p)^{-1}Z_p'y$$
• Two Stages

stage one: OLS of $Q$ on each of the columns of $Z$ yields predicted values $Z_p$

stage two: OLS of $y$ on $Z_p$

proper residuals are $y - Z\theta_{2SLS}$ NOT $y - Z_p\theta_{2SLS}$

$\theta_{2SLS} = [Z'Q(Q'Q)^{-1}Q'Z]^{-1}Z'Q(Q'Q)^{-1}Q'y$
• **Asymptotic Properties**

same as IV estimator with instrument matrix

\[ H = Q(Q'Q)^{-1}Q'Z \]

consistency and asymptotic normality from LLN and CLT, not from assumption of normality

\[
\text{Var}[\theta_{2SLS}] = (H'Z)^{-1}(H'\Sigma H)(Z'H)^{-1} \\
= [Z'Q(Q'Q)^{-1}Q'Z]^{-1} \\
Z'Q(Q'Q)^{-1}(Q'\Sigma Q)(Q'Q)^{-1}Q'Z \\
[Z'Q(Q'Q)^{-1}Q'Z]^{-1}
\]
Illustration
• Homicide Rate Regression - south.dbf

  HR90: county homicide rate

  RD90: resource deprivation

  PS90: population structure component

  MA90: median age

  UE90: unemployment rate
Endogeneity

UE possibly endogenous

instruments:
FP89: % families below poverty rate
GI89: Gini index
FH90: % female headed households
• Ordinary Least Squares Regression

HR90 on RD90, PS90, MA90 and UE90 for reference purposes only

if UE90 is indeed endogenous, then the OLS estimates will be biased
Regression
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION

Data set : south
Dependent Variable : HR90  Number of Observations: 1412
Mean dependent var : 9.54929  Number of Variables : 5
S.D. dependent var : 7.03636  Degrees of Freedom : 1407

R-squared : 0.301265  F-statistic : 151.66
Adjusted R-squared : 0.299278  Prob(F-statistic) : 0
Sum squared residual: 48847.6  Log likelihood : -4505.39
Sigma-square : 34.7175  Akaike info criterion : 9020.78
S.E. of regression : 5.89216  Schwarz criterion : 9047.05
Sigma-square ML : 34.5946
S.E of regression ML: 5.88172

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std.Error</th>
<th>t-Statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>10.5929</td>
<td>1.74352</td>
<td>6.075581</td>
<td>0.00000000</td>
</tr>
<tr>
<td>RD90</td>
<td>4.560765</td>
<td>0.2156085</td>
<td>21.15299</td>
<td>0.00000000</td>
</tr>
<tr>
<td>PS90</td>
<td>2.129831</td>
<td>0.2018352</td>
<td>10.55233</td>
<td>0.00000000</td>
</tr>
<tr>
<td>MA90</td>
<td>-0.002638148</td>
<td>0.04773761</td>
<td>-0.05526351</td>
<td>0.9561890</td>
</tr>
<tr>
<td>UE90</td>
<td>-0.5078603</td>
<td>0.07027898</td>
<td>-7.226348</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

OLS Results
• Explicit Two Stage Least Squares

stage 1: OLS of UE90 on RD90, PS90, MA90 and FP89, GI89 and FH90

stage 2: OLS of HR90 on RD90, PS90, MA90 and predicted value of UE90 from stage 1
Stage 1 with predicted values saved
### Stage 2 Results

**Regression**

**SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION**

<table>
<thead>
<tr>
<th>Data set</th>
<th>south</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td>HR90</td>
</tr>
<tr>
<td>Mean dependent var</td>
<td>9.54929</td>
</tr>
<tr>
<td>S.D. dependent var</td>
<td>7.03636</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>1412</td>
</tr>
<tr>
<td>Number of Variables</td>
<td>5</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>1407</td>
</tr>
</tbody>
</table>

| R-squared | 0.321266 |
| Adjusted R-squared | 0.319337 |
| F-statistic | 166.495 |
| Prob(F-statistic) | 0 |
| Sum squared residual | 47449.3 |
| Log likelihood | -4484.89 |
| Sigma-square | 33.7237 |
| Akaike info criterion | 8979.78 |
| S.E. of regression | 5.80722 |
| Schwarz criterion | 9006.04 |
| Sigma-square ML | 33.6043 |
| S.E of regression ML | 5.79692 |

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std.Error</th>
<th>t-Statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>15.61382</td>
<td>1.869047</td>
<td>8.353892</td>
<td>0.00000000</td>
</tr>
<tr>
<td>RD90</td>
<td>5.965739</td>
<td>0.2957663</td>
<td>20.17045</td>
<td>0.00000000</td>
</tr>
<tr>
<td>PS90</td>
<td>2.261376</td>
<td>0.1998557</td>
<td>11.31504</td>
<td>0.00000000</td>
</tr>
<tr>
<td>MA90</td>
<td>-0.02379913</td>
<td>0.04715132</td>
<td>-0.5047394</td>
<td>0.6138359</td>
</tr>
<tr>
<td>UE90_P</td>
<td>-1.211717</td>
<td>0.1241749</td>
<td>-9.758151</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>
• Proper 2SLS

  same coefficient estimates as two stage approach

• correct residuals

  different error variance

  different coefficient variance estimates and t statistics
2SLS variable selection in GeoDaSpace
### SUMMARY OF OUTPUT: TWO STAGE LEAST SQUARES ESTIMATION

<table>
<thead>
<tr>
<th>Data set</th>
<th>south.dbf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td>HR90</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>1412</td>
</tr>
<tr>
<td>Mean dependent var</td>
<td>9.5493</td>
</tr>
<tr>
<td>Number of Variables</td>
<td>5</td>
</tr>
<tr>
<td>S.D. dependent var</td>
<td>7.0389</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>1407</td>
</tr>
<tr>
<td>Pseudo R-squared</td>
<td>0.268839</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>15.6138196</td>
<td>1.9593407</td>
<td>7.9689152</td>
<td>1.600733e-15</td>
</tr>
<tr>
<td>RD90</td>
<td>5.9657391</td>
<td>0.3100547</td>
<td>19.2409241</td>
<td>1.681831e-82</td>
</tr>
<tr>
<td>PS90</td>
<td>2.2613755</td>
<td>0.2095107</td>
<td>10.7936033</td>
<td>3.690309e-27</td>
</tr>
<tr>
<td>MA90</td>
<td>-0.0237991</td>
<td>0.0494292</td>
<td>-0.4814792</td>
<td>0.6301759</td>
</tr>
<tr>
<td>UE90</td>
<td>-1.2117172</td>
<td>0.1301737</td>
<td>-9.3084614</td>
<td>1.296975e-20</td>
</tr>
</tbody>
</table>

Instruments: FP89, GI89, FH90

---

**2SLS GeoDaSpace Results**
<table>
<thead>
<tr>
<th>Var</th>
<th>OLS</th>
<th>Two Stage</th>
<th>2SLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>10.59</td>
<td>15.61</td>
<td>15.61</td>
</tr>
<tr>
<td>RD90</td>
<td>4.56</td>
<td>5.97</td>
<td>5.97</td>
</tr>
<tr>
<td>PS90</td>
<td>2.13</td>
<td>2.26</td>
<td>2.26</td>
</tr>
<tr>
<td>MA90</td>
<td>-0.003</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>UE90</td>
<td>-0.51</td>
<td>-1.21</td>
<td>-1.21</td>
</tr>
</tbody>
</table>

Comparison: Coefficient Estimates
<table>
<thead>
<tr>
<th>Var</th>
<th>OLS</th>
<th>Two Stage</th>
<th>2SLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.74</td>
<td>1.87</td>
<td>1.96</td>
</tr>
<tr>
<td>RD90</td>
<td>0.22</td>
<td>0.3</td>
<td>0.31</td>
</tr>
<tr>
<td>PS90</td>
<td>0.2</td>
<td>0.2</td>
<td>0.21</td>
</tr>
<tr>
<td>MA90</td>
<td>0.048</td>
<td>0.047</td>
<td>0.049</td>
</tr>
<tr>
<td>UE90</td>
<td>0.07</td>
<td>0.12</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Comparison: Coefficient Standard Errors