5. Spatial regression models

5.1 Basic types of spatial regression models

There are three basic types of spatial regression models which can be chosen subject to the results of the LM and F tests in the standard regression model:
- the spatial cross-regressive model (SLX model),
- the spatial lag model (SAR model),
- the spatial error model (SEM model).

Spatial cross-regressive Model (SLX model)

Substantive spatial dependence can be captured by spatial lags in the explanatory variables X_2, X_3, …, X_k or the endogenous variable Y. In the former case, the spatial lag variables $\mathbf{Wx}_2$, $\mathbf{Wx}_3$, …, $\mathbf{Wx}_k$ will be incorporated into the standard regression model as additional regressors. We term the regression model with spatially lagged exogenous regressors **spatial cross-regressive model (SLX model)**. Substantive spatial interaction can occur in different applications. Output growth of a region may not only depend on own region's initial income but as well on income in adjacent regions. In this case, spillover effects are restricted to neighbourhood regions. Such a restriction may especially hold for spillovers of tacit knowledge which is expected to be exchanged within local areas.

Parameter estimation in the **cross-regressive model (SLX model)** can be performed as in the standard regression model by **OLS**. This results from the fact that spatial lag variables share the properties with the original regressors, which are assumed to be non-stochastic.

Spatial lag model (SAR model)

The **spatial lag model** (SAR model) captures as well substantial spatial dependencies like external effects or spatial interactions. It assumes that such dependencies mani-fests in the spatial lag **Wy** of the dependent variable Y. Regional growth may be fostered by growth in neighbourhood regions by flows of goods for example. In this case, spillover effects are not restricted to adjacent regions but propagated over the entire regional system.

In accordance to the time-series analogue the **pure spatial lag model** is also termed **spatial autoregressive (SAR) model**. In applications the model also incorporates a set of explanatory variables X_1, X_2, …, X_k. This extension is expressed by the term **mixed regressive, spatial autoregressive model**. In all instances OLS estimation will produce biased and inconsistent parameter estimates. We will introduce the **method of instruments (IV method)** and the **method of maximum likelihood (ML method)** as adequate estimation methods for that type of model. Because only the spatial lag **Wy** is relevant for the choice of an alternative estimation method to OLS, the term spatial lag model is often kept in cases where the model is extended by exogeneous X-variables.

Spatial error model (SEM model)

The **spatial error model (SEM model)** is applicable when spatial autocorrelation occurs as nuisance resulting from misspecification or inadequate delineation of spatial units. Unmodelled interaction among regions are restricted to the error terms. In convergence studies, the convergence rate will be properly assessed by stadard estimation methods. However, a random shock occurring in a specific region is not restricted to that region and its neighbourhood but will diffuse across the entire regional system.

Spatial dependence as nuisance entails that the disturbances ε_i are no longer independently identically distributed (i.i.d.), but follow an autoregressive (AR) or moving average (MA) process. In analogy to the Markov process in time-series analysis, the disturbance term is assumed to follow a first order autoregressive (AR) process in the spatial error model. In consideration of the explanation of the dependent variable Y by a set of exogenous variables X_1, X_2, …, X_k, the **spatial error model (SEM)** serves as an abbreviation for a **linear regression model with a spatial autoregressive disturbance**.

In contrast to substantial dependence, spatial dependence in form of nuisance does not entail inconsistency of OLS estimated regression coefficients. However, as their standard errors are biased, significance tests based on OLS estimation can be misleading. In order to allow for valid inference, other estimation principles

must be adopted. We outline **maximum likelihood (ML) estimation** in the **spatial error model**. An alternative is provided by Kelejian and Prucha (1999) in form of a **general moment (GM) estimator** which is not dealt with in our introductory course.

Specification tests

The presence of spatial error autocorrelation can be assessed by the **Moran test** applied on the **residuals of the standard regression model** (see section 4.2**)**. This omnibus test does, however, not point to a basic spatial model. No specific spatial test is available for the **spatial cross-regressive model**. Florax and Folmer (1992) suggest to apply the well-known **F-test for linear restrictions** on the regression coefficients to identify spatial autocorrelation due to omitted lagged exogenous variables. This test requires the estimation of both the restricted and the unrestricted regression model.

For the two other type of spatial models **Lagrange Multiplier (LM) tests** that are tailored for the special spatial settings are available:

- the classic and robust **LM lag test** and
- the classic and robust **LM error test**.

Both type of tets tests have to be performed after estimating the standard regression model (see section 4.2). If the test statistic **LM(lag)** of the **classic test** turns out to be significant, but **LM(error)** is nonsignificant, spatial autocorrelation appears to be substantial, which means that the **spatial lag model** is viewed to be appropriate. In the converse case the **spatial error model** will be a sensible choice for analysing the relationship between Y and the X-variables. When both test statistics, **LM(lag)** and **LM(error)** prove to be significant, the **robust versions LM_{rob}(lag)** and **LM_{rob}(err)** should be applied. While LM(lag) and LM(err) are also affected by spatial error and spatial lag dependence, respectively, the robust LM tests control each for the other effect. If only one type of spatial dependence is present, the robust LM tests are expected to either the SEM or the SAR model. When both robust LM statistics, **LM_{rob}(lag)** and **LM_{rob}(err)**,will be significant, a more complex spatial modelling is desirable. However, in a parsimonious modelling approach, the **original test statistics** with the higher significance, i.e. the lower p-value, could guide the choice of the spatial regression model. According to this rule, in case of

$$p[LM(lag)] < p[LM(err)]$$

the **spatial lag model** will be chosen, whereas for

$$p[LM(err)] < p[LM(lag)]$$

the **spatial error model** will be the preferable spatial setting.

5.2 The spatial cross-regressive model (SLX model)

The **spatial cross-regressive model (SLX model)** presumes that the exploratory variables X_2, X_3, …, X_k as well as their spatial lags LX_2, …, LX_k influence a geo-referenced dependent variable Y. In this approach Y is not only affected by values the variables take in the same region but also they can take in neighbouring regions:

(5.1) $$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + ... + \beta_k\mathbf{x}_k + \gamma_2\mathbf{W}\mathbf{x}_2 + ... + \gamma_k\mathbf{W}\mathbf{x}_k + \boldsymbol{\varepsilon}$$

y is an nx1 vector of the endogenous variable Y, $\mathbf{x}_j$ an nx1 vector of the exogenous variable X_j (where $\mathbf{x}_1$ is a vector of ones for the intercept), **W** an nxn spatial weight matrix and **ε** an nxn vector of disturbances. The parameters β_j j=1, 2,…,k denote the regression coefficients of the exogenous variables X_1, X_2, …, X_k and the parameters γ_j the regression coefficients of the exogeneous spatial lags $\mathbf{W}\mathbf{x}_2$, …, $\mathbf{W}\mathbf{x}_k$. The disturbances ε_i are assumed to meet the standard assumptions for a linear regression model (see section 4.1): expectation of zero, constant variance σ^2 and absence of autocorrelation. For statistical inferences like significance tests we additionally assume the disturbances to be normally distributed.

A more compact form of the spatial cross-regressive model (SLX model) reads

(5.2) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}^*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$.

nxk matrix with observations of the k explanatory variables:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

nx(k-1) matrix with observations of the k-1 lagged explanatory variables:

$$\mathbf{X}^* = \begin{bmatrix} Lx_{12} & \cdots & Lx_{1k} \\ Lx_{22} & \cdots & Lx_{2k} \\ \vdots & \ddots & \vdots \\ Lx_{n2} & \cdots & Lx_{nk} \end{bmatrix}$$

kx1 vector of regression coefficients of exogenous variables:

$$\boldsymbol{\beta} = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_k]'$$

(k-1)x1 vector of regression coefficients of lagged exogenous variables:

$$\boldsymbol{\gamma} = [\gamma_2 \quad \cdots \quad \gamma_k]'$$

OLS estimator of the regression coefficients:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = ([\mathbf{X} \quad \mathbf{X}^*]'[\mathbf{X} \quad \mathbf{X}^*])^{-1}[\mathbf{X} \quad \mathbf{X}^*]'\mathbf{y}$$

Variance-covariance matrix of OLS estimator $[\hat{\boldsymbol{\beta}} \quad \hat{\boldsymbol{\gamma}}]'$:

$$\mathrm{Cov}\left([\hat{\boldsymbol{\beta}} \quad \hat{\boldsymbol{\gamma}}]'\right) = \sigma^2 \cdot ([\mathbf{X} \quad \mathbf{X}^*]'[\mathbf{X} \quad \mathbf{X}^*])^{-1}$$

- **F test on omitted spatially lagged exogenous variables**

Test of the spatially lagged variables for a relevant subset S of X-variables one at a time

Null hypothesis H_0: $\gamma_j = 0$ for $X_j \in S$

SSR_c: Constrained residual sum of squares from a regression in which H_0 holds i.e. a regression of Y on the original exogenous variables $X_1, X_2, \ldots, X_k$

SSR_u: Unconstrained residual sum of squares from a regression of Y on the original exogenous variables $X_1, X_2, \ldots, X_k$ and the spatially lagged exogenous variable LX_j

Test statistic:

(5.3) $$F = \frac{(SSR_c - SSR_u)}{SSR_u/(n-k-1)}$$

F follows an F distribution with 1 and n-k-1 degrees of freedom.

Testing decision: $F > F(1;n-k-1;1-\alpha)$ => Reject H_0

or

$p < \alpha$ => Reject H_0

Generalisation of the F test:

Instead of testing the spatially lagged variables of the subset S one at a time, they can also be tested simultanously

Null hypothesis H_0: All regression coefficients γ_j of the X-variables of the subset S are equal to zero

Test statistic:

(5.4) $$F = \frac{(SSR_c - SSR_u)/q}{SSR_u/(n-k-q)}$$

q: number of X-variables in the subset S (when all X-variables without the vector of one are included in S, q is equal to k – 1)

F follows an F distribution with q and n-k-q degrees of freedom.

Testing decision: $F > F(q;n-k-q;1-\alpha)$ => Reject H_0

or

$p < \alpha$ => Reject H_0

Example:
In order to illustrate the spatial cross-regressive model (SLX model) , we refer to the example of 5 regions for which data are available on output growth (X) and productivity growth (Y):

Region	1	2	3	4	5
Output growth (X)	0.6	1.0	1.6	2.6	2.2
Productivity growth (Y)	0.4	0.6	0.9	1.1	1.2

The spatial cross-regressive model (SLX model) presumes that productivity growth in region i does not only depend on own region‘s output growth but as well on output growthin adjacent regions

(5.5) $$y_i = \beta_1 + \beta_2 \cdot x_i + \gamma_2 \cdot \sum_{j=1}^{n} w_{ij} \cdot x_j + \varepsilon_i$$

with x_{i1}=1 for all i and $x_{i2} = x_i$. When LX captures technological externalities from neighbouring regions γ_2 is expected to take a positive sign.

When output growth X can be treated as an exogenous variable, the spatial lag variable LX is as well exogenous, because the spatial weights are determined a priori. Thus, the spatial regression model (5.5) can be estimated by OLS.

Vector of the endogenous variable **y**: $y = \begin{bmatrix} 0.4 & 0.6 & 0.9 & 1.1 & 1.2 \end{bmatrix}'$

Matrix of original x-variables **X**:

$$\mathbf{X} = \begin{bmatrix} 1 & 0.6 \\ 1 & 1.0 \\ 1 & 1.6 \\ 1 & 2.6 \\ 1 & 2.2 \end{bmatrix}$$

Standardized weights matrix **W**:

$$\mathbf{W} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Matrix of spatially lagged exogenous variables **X*** :

$$\mathbf{X}^* = \mathbf{W} \cdot \mathbf{x}_2 = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.6 \\ 1.0 \\ 1.6 \\ 2.6 \\ 2.2 \end{bmatrix} = \begin{bmatrix} 1.3 \\ 1.6 \\ 1.4 \\ 1.6 \\ 2.6 \end{bmatrix}$$

Observation matrix [**X** **X***]:

$$\begin{bmatrix} \mathbf{X} & \mathbf{X}^* \end{bmatrix} = \begin{bmatrix} 1 & 0.6 & 1.3 \\ 1 & 1.0 & 1.6 \\ 1 & 1.6 & 1.4 \\ 1 & 2.6 & 1.6 \\ 1 & 2.2 & 2.6 \end{bmatrix}$$

Matrix product [**X X*]'[X X*]**:

$$[\mathbf{X}\ \mathbf{X}^*]'[\mathbf{X}\ \mathbf{X}^*] = \begin{bmatrix} 5 & 8 & 8.5 \\ 8 & 15.52 & 14.5 \\ 8.5 & 14.5 & 15.53 \end{bmatrix}$$

Matrix product [**X X*]'y**:

$$[\mathbf{X}\ \mathbf{X}^*]'\mathbf{y} = \begin{bmatrix} 4.2 \\ 7.78 \\ 7.62 \end{bmatrix}$$

Inverse ([**X X*]'[X X*]**)$^{-1}$:

$$([\mathbf{X}\ \mathbf{X}^*]'[\mathbf{X}\ \mathbf{X}^*])^{-1} = \begin{bmatrix} 2.8930 & -0.0931 & -1.4965 \\ -0.0931 & 0.5076 & -0.4230 \\ -1.4965 & -0.4230 & 1.2784 \end{bmatrix}$$

OLS estimator of the regression coefficients:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = ([\mathbf{X}\ \mathbf{X}^*]'[\mathbf{X}\ \mathbf{X}^*])^{-1}[\mathbf{X}\ \mathbf{X}^*]'\mathbf{y}$$

$$= \begin{bmatrix} 2.8930 & -0.0931 & -1.4965 \\ -0.0931 & 0.5076 & -0.4230 \\ -1.4965 & -0.4230 & 1.2784 \end{bmatrix} \begin{bmatrix} 4.2 \\ 7.78 \\ 7.62 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 = 0.0230 \\ \hat{\beta}_2 = 0.3350 \\ \hat{\gamma}_2 = 0.1653 \end{bmatrix}$$

Vector of fitted values $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = [\mathbf{X} \;\; \mathbf{X}^*] \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} 1 & 0.6 & 1.3 \\ 1 & 1.0 & 1.6 \\ 1 & 1.6 & 1.4 \\ 1 & 2.6 & 1.6 \\ 1 & 2.2 & 2.6 \end{bmatrix} \begin{bmatrix} 0.0230 \\ 0.3350 \\ 0.1653 \end{bmatrix} = \begin{bmatrix} 0.4389 \\ 0.6225 \\ 0.7904 \\ 1.1585 \\ 1.1897 \end{bmatrix}$$

Vector of residuals **e**:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 0.4 \\ 0.6 \\ 0.9 \\ 1.1 \\ 1.2 \end{bmatrix} - \begin{bmatrix} 0.4389 \\ 0.6225 \\ 0.7904 \\ 1.1585 \\ 1.1897 \end{bmatrix} = \begin{bmatrix} -0.0389 \\ -0.0225 \\ 0.1096 \\ -0.0585 \\ 0.0103 \end{bmatrix}$$

Residual variance $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{5-3} = \frac{1}{2}\begin{bmatrix} -0.0389 & -0.0225 & 0.1096 & -0.0585 & 0.0103 \end{bmatrix}\begin{bmatrix} -0.0389 \\ -0.0225 \\ 0.1096 \\ -0.0585 \\ 0.0103 \end{bmatrix}$$

$$= \frac{0.0176}{2} = 0.0088$$

Standard error of regression (SER):

$$\text{SER} = \sqrt{0.0088} = 0.0937$$

Estimated variance-covariance matrix of $[\hat{\boldsymbol{\beta}} \quad \hat{\gamma}]'$:

$$\hat{\text{Cov}}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} & \hat{\boldsymbol{\gamma}} \end{bmatrix}'\right) = \hat{\sigma}^2 \cdot \left(\begin{bmatrix} \mathbf{X} & \mathbf{X}* \end{bmatrix}'\begin{bmatrix} \mathbf{X} & \mathbf{X}* \end{bmatrix}\right)^{-1}$$

$$= 0.0088 \cdot \begin{bmatrix} 2.8930 & -0.0931 & -1.4965 \\ -0.0931 & 0.5076 & -0.4230 \\ -1.4965 & -0.4230 & 1.2784 \end{bmatrix}$$

Coefficient of determination

Working table ($\bar{y} = 0.84$)

i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	0.4	0.4389	-0.44	0.1936	-0.4011	0.1609
2	0.6	0.6225	-0.24	0.0576	-0.2175	0.0473
3	0.9	0.7904	0.06	0.0036	-0.0496	0.0025
4	1.1	1.1585	0.26	0.0676	0.3185	0.1014
5	1.2	1.1897	0.36	0.1296	0.3497	0.1223
Σ	4.2	4.2	0	0.4520	0	0.4344

SST = 0.4520, SSE = 0.4344, SSR = SST – SSE = 0.4520 -0.4344 = 0.0176

$$R^2 = \frac{SSE}{SST} = \frac{0.4344}{0.4520} = 0.961$$ or

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{0.0176}{0.4520} = 1 - 0.039 = 0.961$$

Test of significance of regression coefficients

- for β_1 (H_0: $\beta_1 = 0$)

OLS estimator for β_1: $\hat{\beta}_1 = 0.0230$

Test statistic: $t_{\beta_1} = \frac{\hat{\beta}_1}{\hat{\sigma} \cdot \sqrt{xx^{11}}} = \frac{0.0230}{0.0937 \cdot \sqrt{2.8930}} = 0.144$

Critical value (α=0.05, two-sided test): t(2,0.975) = 4.303

Testing decision: (| t_1 |=0.144) < [t(2;0.975)=4.303] => Accept H_0

- for β_2 (H_0: $\beta_2 = 0$)

OLS estimator for β_2: $\hat{\beta}_2 = 0.3350$

Test statistic: $t_{\beta_2} = \frac{\hat{\beta}_2}{\hat{\sigma} \cdot \sqrt{xx^{22}}} = \frac{0.3350}{0.0937 \cdot \sqrt{0.5076}} = 5.018$

Critical value (α=0.05, two-sided test): t(2,0.975) = 4.303

Testing decision: (| t_2 |=5.018) > [t(2;0.975)=4.303] => Reject H_0

Test of significance of regression coefficients

- for γ_2 (H_0: $\gamma_2 = 0$)

OLS estimator for γ_2: $\hat{\gamma}_1 = 0.1653$

Test statistic: $t_{\gamma_2} = \frac{\hat{\gamma}_2}{\hat{\sigma} \cdot \sqrt{x^* x^{*11}}} = \frac{0.1653}{0.0937 \cdot \sqrt{1.2784}} = 1.560$

Critical value (α=0.05, two-sided test): t(2,0.975) = 4.303

Testing decision: (| t_1 |=1.560) < [t(2;0.975)=4.303] => Accept H_0

F test for the regression as a whole

Null hypothesis H_0: $\beta_2 = \gamma_2 = 0$

Constrained residual sum of squares: $SSR_c = SST = 0.4520$
Unconstrained residual sum of squares: $SSR_u = SSR = 0.0176$

Test statistic:

$$F = \frac{(SSR_c - SSR_u)/k}{SSR_u/(n-k-1)} = \frac{(0.4520 - 0.0176)/2}{0.0176/(5-2-1)} = 24.682$$

or $$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.961/2}{(1-0.961)/(5-2-1)} = 24.641$$

(The difference of both computations of F are only due to rounding errors.)

Critical value(α=0.05): F(2;2;0.95) = 19.0

Testing decision: (F=24.682) > [F(2;2;0.95)=19.0] => Reject H_0

F test on omitted spatially lagged exogenous variables

Null hypothesis H_0: $\gamma_2 = 0$

Constrained residual sum of squares (regression of Y on X_1 (=constant) and X_2): $SSR_c = 0.0388$ (see sect. 4.1 → standard regression model)

Unconstrained residual sum of squares (regression of Y on X_1 (constant), X_2 and LX_2): $SSR_u = 0.0176$

Test statistic:

$$F = \frac{(SSR_c - SSR_u)}{SSR_u/(n-k-1)} = \frac{(0.0388 - 0.0176)}{0.0176/(5-2-1)} = 2.409$$

Critical value(α=0.05): F(1;2;0.95) = 18.5

Testing decision: (F=2.409) < [F(1;2;0.95)=18.5] => Accept H_0