4. Standard Regression Model and Spatial Dependence Tests

Standard regression analysis fails in the presence of spatial effects. In case of spatial dependencies and/or spatial heterogeneity a standard regression model will be misspecified. Spatial effects have to be incorporated in regression models in order to obtain valid parameter estimates. Here we focus on spatial dependence of which ignorance causes severe interpretation problems and requires an original spatial modelling approach. Spatial heterogeneity can much more be accounted for by methods developed in mainstream econometrics.

The standard regression model is usually the starting point of spatial regression analysis. The residuals of ordinary least-squares (OLS) estimation can be used to test for spatial effects. Hence, we first outline OLS estimation in the standard regression model in section 4.1. In section 4.2 we introduce various test for spatial effects in the regression model. Spatial regression models suggested by the outcomes of the tests are presented in chapter 5.
4.1 The standard regression model

Linear regression model:
Relationship between a dependent variable Y and a set of explanatory variables $X_1, X_2, \ldots, X_k$.

\[(4.1) \quad y = X \cdot \beta + \varepsilon\]

nx1 vector of the dependent variable: $y = [y_1 \quad y_2 \ldots y_n]'$

nxk matrix with observations of the k explanatory variables:

\[
X = \begin{bmatrix}
1 & x_{12} & \ldots & x_{1k} \\
1 & x_{22} & \ldots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n2} & \ldots & x_{nk}
\end{bmatrix}
\]

$x_{ij}$: observation of the $j^{th}$ variable at the $i^{th}$ statistical unit
1st column of $X$: vector of ones (for intercept)

The explanatory variables are treated as fixed and not random.

kx1 vector of regression coefficients: $\beta = [\beta_1 \quad \beta_2 \ldots \beta_k]'$

nx1 vector of disturbances (error terms): $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \ldots \varepsilon_n]'$
• Standard assumptions:

1. The disturbance has an expectation of zero:
   
   \[ E(\varepsilon_i) = 0 \quad \text{for all } i \]

2. The disturbances have a constant variance (homoscedasticity):
   
   \[ \operatorname{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2 \quad \text{for all } i, \quad \sigma^2: \text{error variance} \]

3. The disturbances are uncorrelated (lack of autocorrelation):
   
   \[ \operatorname{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) = 0 \quad \text{for all } i \neq j \]

Assumptions 1-3 in compact form:

\[ E(\varepsilon) = \mathbf{0} \quad \text{and} \quad \operatorname{Cov}(\varepsilon) = E(\varepsilon \cdot \varepsilon^\prime) = \sigma^2 \cdot \mathbf{I} \]

\( \mathbf{0} \): nx1 vector of zeros, \( \mathbf{I} \): nxn identity matrix

For carrying out statistical tests normality of the errors is assumed:

\[ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{for all } i \quad \text{or} \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \]
• Ordinary least squares (OLS) estimation

An important task of regression analysis is to estimate the unknown vector of regression coefficients, $\beta$, in order to assess the influence of the regressors $X_1, X_2, \ldots, X_k$ on the dependent variable $Y$. Under the standard assumptions, ordinary least squares (OLS) estimation yields best linear unbiased estimators (blue property).

Least squares criterion:

$$(4.2a) \quad Q(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (y - X\beta)'(y - X\beta).$$

$Q$ has to be minimized with respect to $\beta$ for which we use the equivalent expression:

$$(4.2b) \quad Q(\beta) = y'y - 2y'X\beta + \beta'X'X\beta$$

First order condition for a minimum of $Q$:

$$\frac{dQ(\beta)}{d\beta} = -2X'y + 2(X'X)\beta = 0$$

OLS estimator of $\beta$;

$$(4.3) \quad \hat{\beta} = (X'X)^{-1}X'y$$
• Fitted values, residuals and residual variance

Fitted values: (4.4) \( \hat{y} = X \cdot \hat{\beta} \)

Residuals: (4.5a) \( e = y - \hat{y} \) or (4.5b) \( e = y - X \cdot \hat{\beta} \)

Residual variance (unbiased estimate of \( \sigma^2 \)):

(4.6) \( \hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} e_i^2 = \frac{e'e}{n-k} \quad (\bar{e} = 0) \)

Standard error of regression (SER):

(4.7) \( \text{SER} = \hat{\sigma} = \sqrt{\hat{\sigma}^2} \)
• Measures of fit

Decomposition of the total sum of squares of the dependent variable $Y$:

(4.8) $SST = SSE + SSR$

Total sum of squares:

(4.9) $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Explained sum of squares:

(4.10) $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Residual sum of squares:

(4.11) $SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e$

Coefficient of determination:

(4.12a) $R^2 = \frac{SSE}{SST}$ or (4.12b) $R^2 = 1 - \frac{SSR}{SST}$

Range of $R^2$: $0 \leq R^2 \leq 1$
Adjusted coefficient of determination:
By the adjustment regression models with different numbers of regressors are
made comparable.

\[
(4.13) \quad \overline{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2)
\]

Information criteria

Information measure the goodness of fit where model complexity in terms of
the number of explanatory variables is penalized. Goodness of fit is covered
by the log likelihood function \( \ln L \),

\[
(4.14) \quad \ln L = C - \frac{n}{2} \cdot \ln(\hat{\sigma}^2),
\]

which is mainly composed of the sum of squared residuals. By penalizing fits
with a larger number of regressors, regression models with different \( k \) are
made comparable. According to the information criteria, the model with the
lowest value is the best.

- Akaike information criterion (AIC)

\[
(4.15a) \quad \text{AIC} = -2 \cdot \ln(L) + 2k
\]

- Schwartz criterion (SC)

\[
(4.15b) \quad \text{SC} = -2 \cdot \ln(L) + k \cdot \ln(n)
\]
- Hypothesis tests

- Test of significance of regression coefficients

Null hypothesis $H_0$: $\beta_j = 0$

Distribution of the OLS estimator $\hat{\beta}$ under $H_0$ for normally distributed errors:

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

because of $E(\hat{\beta}) = \beta$ and $Cov(\hat{\beta}) = \sigma^2 \cdot (X'X)^{-1}$

Test statistic:

$$(4.16) \quad t_j = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{xx_{jj}}}$$

$xx_{jj}$: $j$th main diagonal element of the inverse $(X'X)^{-1}$

t$_j$ follows a t distribution with $n-k$ degrees of freedom.

Significance level: $\alpha$

Critical value (two-sided test): $t(n-k;1-\alpha/2)$

Testing decision: $|t_j| > t(n-k;1-\alpha/2) \Rightarrow$ Reject $H_0$

or

$$p < \alpha \Rightarrow$ Reject $H_0$

p-value: Probability of obtaining an (absolute) higher t statistic than $|t_j|$
- F test for the regression as a whole

Null hypothesis $H_0: \beta_2 = \beta_3 = \ldots = \beta_k = 0$

$SSR_c$: Constrained residual sum of squares from a regression in which $H_0$ holds i.e. a regression of $Y$ on the constant term $X_1$ only

$SSR_u$: Unconstrained residual sum of squares from a regression of $Y$ on $X_1, X_2, \ldots, X_k$

Test statistic:

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

or

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

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

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

or

$p < \alpha \Rightarrow$ Reject $H_0$
Example:

For 5 regions are data available on output growth (X) and productivity growth (Y):

<table>
<thead>
<tr>
<th>Region</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output growth (X)</td>
<td>0.6</td>
<td>1.0</td>
<td>1.6</td>
<td>2.6</td>
<td>2.2</td>
</tr>
<tr>
<td>Productivity growth (Y)</td>
<td>0.4</td>
<td>0.6</td>
<td>0.9</td>
<td>1.1</td>
<td>1.2</td>
</tr>
</tbody>
</table>

According to the Verdoorn law output growth and productivity growth are positively related. Productivity growth increases with output growth due to increasing returns to scale. The regression model implied by Verdoorn’s law reads

\[(4.18) \quad y_i = \beta_1 + \beta_2 \cdot x_i + \varepsilon_i\]

with \(x_{i1} = 1\) for all \(i\) and \(x_{i2} = x_i\). If Verdoorn’s law holds, the Verdoorn coefficient \(\beta_2\) is expected to take a positive sign. The intercept captures productivity growth evoked by autonomous technical progress.

The regression model (4.17) can be estimated by OLS.
Vector of the endogenous variable $y$:

$$y = [0.4 \quad 0.6 \quad 0.9 \quad 1.1 \quad 1.2]'$$

Observation matrix $X$:

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

Matrix product $X'X$, its inverse $(X'X)^{-1}$, matrix product $X'y$:

$$X'X = \begin{bmatrix}
5 & 8 \\
8 & 15.52 \\
\end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix}
1.1412 & -0.5882 \\
-0.5882 & 0.3676 \\
\end{bmatrix}, \quad X'y = \begin{bmatrix}
4.20 \\
7.78 \\
\end{bmatrix}$$

OLS estimator of $\beta$:

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix}
1.1412 & -0.5882 \\
-0.5882 & 0.3676 \\
\end{bmatrix} \begin{bmatrix}
4.20 \\
7.78 \\
\end{bmatrix} = \begin{bmatrix}
0.2165 \\
0.3897 \\
\end{bmatrix}$$
Vector of fitted values $\hat{y}$:

$$\hat{y} = X\hat{\beta} = \begin{bmatrix} 1 & 0.6 \\ 1 & 1.0 \\ 1 & 1.6 \\ 1 & 2.6 \\ 1 & 2.2 \end{bmatrix} \begin{bmatrix} 0.2165 \\ 0.3897 \end{bmatrix} = \begin{bmatrix} 0.4503 \\ 0.6062 \\ 0.8400 \\ 1.2297 \\ 1.0738 \end{bmatrix}$$

Vector of residuals $e$:

$$e = y - \hat{y} = \begin{bmatrix} 0.4 \\ 0.6 \\ 0.9 \\ 1.1 \\ 1.2 \end{bmatrix} - \begin{bmatrix} 0.4503 \\ 0.6062 \\ 0.8400 \\ 1.2297 \\ 1.0738 \end{bmatrix} = \begin{bmatrix} -0.0503 \\ -0.0062 \\ 0.0600 \\ -0.1297 \\ 0.1262 \end{bmatrix}$$
Residual variance $\hat{\sigma}^2$:

$$
\hat{\sigma}^2 = \frac{e'e}{5 - 2} = \frac{1}{3} \begin{bmatrix}
-0.0503 & -0.0062 & 0.0600 & -0.1297 & 0.1262 \\
-0.0062 & 0.0600 & 0.1297 & -0.1262 & \\
0.0600 & 0.1297 & -0.1262 & \\
-0.1297 & -0.1262 & \\
0.1262 & \\
\end{bmatrix}
$$

$$
= \frac{0.0389}{3} = 0.0130
$$

Standard error of regression (SER):

$$
SER = \sqrt{0.0130} = 0.1139
$$
Coefficient of determination

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

<table>
<thead>
<tr>
<th>i</th>
<th>$y_i$</th>
<th>$\hat{y}_i$</th>
<th>$y_i - \bar{y}$</th>
<th>$(y_i - \bar{y})^2$</th>
<th>$\hat{y}_i - \bar{y}$</th>
<th>$(\hat{y}_i - \bar{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.4503</td>
<td>-0.44</td>
<td>0.1936</td>
<td>-0.3897</td>
<td>0.1519</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>0.6062</td>
<td>-0.24</td>
<td>0.0576</td>
<td>-0.2338</td>
<td>0.0547</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>0.8400</td>
<td>0.06</td>
<td>0.0036</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>1.2297</td>
<td>0.26</td>
<td>0.0676</td>
<td>0.3897</td>
<td>0.1519</td>
</tr>
<tr>
<td>5</td>
<td>1.2</td>
<td>1.0738</td>
<td>0.36</td>
<td>0.1296</td>
<td>0.2338</td>
<td>0.0547</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>4.2</td>
<td>4.2</td>
<td>0</td>
<td>0.4520</td>
<td>0</td>
<td>0.4132</td>
</tr>
</tbody>
</table>

$SST = 0.4520$, $SSE = 0.4132$, $SSR = SST - SSE = 0.4520 - 0.4132 = 0.0388$

$$R^2 = \frac{SSE}{SST} = \frac{0.4132}{0.4520} = 0.914$$

or

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{0.0388}{0.4520} = 1 - 0.086 = 0.914$$
Test of significance of regression coefficients

- for $\beta_1$  ($H_0: \beta_1 = 0$)

  OLS estimator for $\beta_1$: $\hat{\beta}_1 = 0.2165$

  Test statistic: $t_1 = \frac{\hat{\beta}_1}{\hat{\sigma} \cdot \sqrt{xx^{11}}} = \frac{0.2165}{0.1139 \cdot \sqrt{1.1412}} = 1.779$

  Critical value ($\alpha=0.05$, two-sided test): $t(3,0.975) = 3.182$

  Testing decision: ($|t_1|=1.779 < [t(3;0.975)=3.182]$) $\Rightarrow$ Accept $H_0$

- for $\beta_2$  ($H_0: \beta_2 = 0$)

  OLS estimator for $\beta_2$: $\hat{\beta}_2 = 0.3897$

  Test statistic: $t_2 = \frac{\hat{\beta}_2}{\hat{\sigma} \cdot \sqrt{xx^{22}}} = \frac{0.3897}{0.1139 \cdot \sqrt{0.3676}} = 5.643$

  Critical value ($\alpha=0.05$, two-sided test): $t(3,0.975) = 3.182$

  Testing decision: ($|t_2|=5.643 > [t(3;0.975)=3.182]$) $\Rightarrow$ Reject $H_0$
F test for the regression as a whole

Null hypothesis $H_0$: $\beta_2 = 0$ (only one non constant exogenous variable)

Constrained residual sum of squares: $SSR_c = SST = 0.4520$
Unconstrained residual sum of squares: $SSR_u = SSR = 0.0388$

Test statistic:
\[
F = \frac{(SSR_c - SSR_u) / (k - 1)}{SSR_u / (n - k)} = \frac{(0.4520 - 0.0388) / (2 - 1)}{0.0388 / (5 - 2)} = 31.948
\]

or
\[
F = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)} = \frac{0.914 / (2 - 1)}{(1 - 0.914) / (5 - 2)} = 31.884
\]

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

Critical value($\alpha=0.05$): $F(1;3;0.95) = 10.1$

Testing decision: $(F=31.948) > [F(1;3;0.95)=10.1]$  =>  Reject $H_0$