Assignment 8 Solutions: Spatial Data II

Part 2: Spatial Lag Model and Model Comparison

Applied Quantitative Methods II, UC3M

Spring 2026

Contents


```
library(sf)
library(spData)
library(spdep)
library(spatialreg)
library(ggplot2)

data(world)
```

1. Spatial Lag Model (SLM)

Fit the Spatial Lag Model (SAR) using `lagsarlm()`:

```
slm_fit = lagsarlm(lifeExp ~ log_gdp, data = world,
                   listw = listw, zero.policy = TRUE)
summary(slm_fit)
```

```
##
## Call:lagsarlm(formula = lifeExp ~ log_gdp, data = world, listw = listw,
##     zero.policy = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.4987  -2.3164   1.0183   3.2982   8.1470
##
## Type: lag
## Regions with no neighbours included:
##  1 18 39 40 71 82 88 125 126 127 128 133 136 143 147 159
```

```
## Coefficients: (asymptotic standard errors)
##             Estimate Std. Error z value             Pr(>|z|)
## (Intercept) 20.46470    3.11450  6.5708      0.00000000005005
## log_gdp      5.54820    0.32416 17.1157 < 0.00000000000000022
##
## Rho: -0.0042561, LR test value: 0.060951, p-value: 0.805
## Asymptotic standard error: 0.017336
##     z-value: -0.2455, p-value: 0.80607
## Wald statistic: 0.060271, p-value: 0.80607
##
## Log likelihood: -479.9635 for lag model
## ML residual variance (sigma squared): 23.61, (sigma: 4.859)
## Number of observations: 160
## Number of parameters estimated: 4
## AIC: NA (not available for weighted model), (AIC for lm: 965.99)
## LM test for residual autocorrelation
## test value: 54.825, p-value: 0.00000000000013178
```

a) The $\hat{\rho}$ (rho) parameter and its p-value are shown in the summary above, along with the coefficient on `log_gdp`. If $\hat{\rho}$ is positive and statistically significant ($p < 0.05$), it confirms that a country's life expectancy is partly a function of its neighbors' life expectancy.

b) $\hat{\rho}$ captures the degree of spatial diffusion in the outcome itself. A positive and significant $\hat{\rho}$ means that countries surrounded by neighbors with high life expectancy tend to have higher life expectancy themselves, and vice versa — over and above what their own GDP level predicts. This is consistent with cross-border spillovers: regional health systems, shared disease environments, migration patterns, and common institutions create geographic clusters of similar life expectancy that a purely domestic predictor like GDP cannot capture alone.

c) In the SLM, solving for $\mathbf{y}$ yields $\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$. Because of the inverse matrix $(\mathbf{I} - \rho\mathbf{W})^{-1}$, a change in x_i does not affect only y_i: it ripples through the network to all other units via the spatial feedback loop. The raw `log_gdp` coefficient in the SLM output captures only the **direct, first-order** effect before this feedback is accounted for; the true marginal effect is found by computing the equilibrium impacts using `impacts()`.

2. Direct and Indirect Effects

a) Compute equilibrium direct, indirect, and total effects via simulation:

```
set.seed(42)
impacts_slm = impacts(slm_fit, listw = listw, R = 500)
summary(impacts_slm, zstats = TRUE)
```

```
## Impact measures (lag, exact):
##           Direct    Indirect    Total
## log_gdp 5.548223 -0.02353893 5.524685
## ========================================================
## Simulation results ( variance matrix):
## Direct:
##
## Iterations = 1:500
```

```
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 500
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean     SD Naive SE Time-series SE
## log_gdp 5.553 0.3105  0.01389        0.01389
##
## 2. Quantiles for each variable:
##
##          2.5%  25%   50%   75% 97.5%
## log_gdp 4.952 5.36 5.562 5.753 6.127
##
## ============================================================
## Indirect:
##
## Iterations = 1:500
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 500
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##              Mean      SD Naive SE Time-series SE
## log_gdp -0.01722 0.09542 0.004267       0.004267
##
## 2. Quantiles for each variable:
##
##            2.5%      25%      50%     75% 97.5%
## log_gdp -0.1971 -0.07743 -0.01817 0.04168  0.18
##
## ============================================================
## Total:
##
## Iterations = 1:500
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 500
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean     SD Naive SE Time-series SE
## log_gdp 5.535 0.3205  0.01434        0.01434
```

```
##
## 2. Quantiles for each variable:
##
##          2.5%   25%   50%   75% 97.5%
## log_gdp 4.917 5.327 5.542 5.746 6.166
##
## ======================================================
## Simulated standard errors
##            Direct   Indirect     Total
## log_gdp 0.3105482 0.09541923 0.3205414
##
## Simulated z-values:
##           Direct   Indirect    Total
## log_gdp 17.88032 -0.1804497 17.26917
##
## Simulated p-values:
##         Direct                  Indirect Total
## log_gdp < 0.00000000000000222 0.8568   < 0.00000000000000222
```

The `impacts()` output reports three quantities for `log_gdp`: the **direct effect** (average effect of a country's own log GDP on its own life expectancy, accounting for feedback), the **indirect effect** (average spillover to all other countries' life expectancy), and the **total effect** (their sum). The direct effect is slightly larger than the raw SLM coefficient because it includes own-feedback from the equilibrium matrix. Compared to the OLS coefficient, the total effect is larger because it compounds the direct effect with the network-propagated spillovers that OLS ignores entirely.

b) The indirect effect measures the average change in life expectancy across all other countries that results from a 1-unit increase in log GDP per capita in one focal country — after the spatial feedback loop has reached equilibrium. If Country A's log GDP per capita rises by 1 unit, the indirect effect tells us by how much neighboring (and more distant, discounted by network distance) countries' life expectancy changes on average. The magnitude reflects both the strength of $\hat{\rho}$ and the density of the neighborhood graph.

c) The total effect exceeding the direct effect is an expected feature of the SLM whenever $\hat{\rho} > 0$. As $\rho \to 0$ the equilibrium matrix approaches the identity matrix, the feedback loop vanishes, and the indirect effect approaches zero — leaving total = direct = the raw coefficient. As ρ grows larger, the feedback multiplier $(\mathbf{I} - \rho\mathbf{W})^{-1}$ amplifies shocks further through the network, producing larger indirect effects and a bigger wedge between total and direct. The indirect effect is therefore larger when both ρ is large and countries are densely connected in the weights graph.

3. Model Comparison

a) Compare OLS, SEM, and SLM by AIC:

```
AIC(ols_fit)
```

```
## [1] 965.988
```

```
AIC(sem_fit)
```

```
## [1] 894.7021
```

```
AIC(slm_fit)
```

```
## [1] 967.927
```

The three AIC values are shown above. A lower AIC indicates better fit penalized for model complexity. The spatial models (SEM and SLM) both improve on OLS, confirming that accounting for spatial dependence is warranted. Whichever spatial model achieves the lower AIC is consistent with the LM-test-based decision from Part 1 (question 1.3b): the robust LM tests and the AIC should point to the same preferred model, providing a useful cross-check between specification tests and information criteria.

b) Summary: The Moran's I test on OLS residuals revealed strong and statistically significant positive spatial autocorrelation ($I > 0$, $p < 0.001$), meaning the OLS independence-of-errors assumption is violated. The LM tests showed that both SEM and SLM are candidates; the robust versions indicated which model fits better after controlling for the other form of dependence. The key `log_gdp` coefficient is broadly similar across OLS, SEM, and SLM, but the SEM coefficient is corrected for spatial confounding in the error term while the SLM direct effect accounts for network feedback. The SLM further reveals that life expectancy spillovers across borders are real and non-trivial in magnitude, as reflected in the positive and significant $\hat{\rho}$ and the indirect effect from `impacts()`. One limitation of queen contiguity weights for country-level data is that island nations have no neighbors and are excluded from the spatial structure; moreover, contiguity ignores the strength of actual cross-border linkages (trade, migration, shared institutions), treating all shared-border pairs as equally connected regardless of geographic or economic distance.

4. Extension: Spatial Durbin Model (optional/bonus)

```
sdm_fit = lagsarlm(lifeExp ~ log_gdp, data = world,
                   listw = listw, Durbin = TRUE, zero.policy = TRUE)
summary(sdm_fit)
```

```
##
## Call:lagsarlm(formula = lifeExp ~ log_gdp, data = world, listw = listw,
##     Durbin = TRUE, zero.policy = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17.32228  -2.12897   0.83353   2.63920   7.57252
##
## Type: mixed
## Regions with no neighbours included:
##  1 18 39 40 71 82 88 125 126 127 128 133 136 143 147 159
## Coefficients: (asymptotic standard errors)
##             Estimate Std. Error z value             Pr(>|z|)
## (Intercept) 17.11312    3.07966  5.5568      0.00000002747353
## log_gdp      5.98747    0.32355 18.5054 < 0.00000000000000022
## lag.log_gdp -3.82746    0.54995 -6.9597      0.00000000000341
##
## Rho: 0.48091, LR test value: 28.767, p-value: 0.000000081634
## Asymptotic standard error: 0.071225
##     z-value: 6.752, p-value: 0.000000000014585
## Wald statistic: 45.589, p-value: 0.000000000014585
##
```

```
## Log likelihood: -464.9167 for mixed model
## ML residual variance (sigma squared): 18.283, (sigma: 4.2758)
## Number of observations: 160
## Number of parameters estimated: 5
## AIC: NA (not available for weighted model), (AIC for lm: 966.6)
## LM test for residual autocorrelation
## test value: 27.91, p-value: 0.00000012711
```

a) The SDM adds `lag.log_gdp` (the spatially lagged version of log GDP per capita, $\mathbf{W} \times$ log_gdp) alongside the own-country `log_gdp` and the spatially lagged dependent variable. If the `lag.log_gdp` coefficient is statistically significant, it means that a neighbor's GDP per capita predicts a country's life expectancy beyond what the country's own GDP already explains — a direct spillover from neighbors' economic conditions to own health outcomes, separate from the outcome-diffusion channel captured by $\hat{\rho}$.

b) Compare AIC of SDM to SEM and SLM:

```
AIC(sdm_fit)
```

```
## [1] 939.8333
```

```
AIC(sem_fit)
```

```
## [1] 894.7021
```

```
AIC(slm_fit)
```

```
## [1] 967.927
```

The SDM has one additional parameter (`lag.log_gdp`) compared to the SLM. If the SDM's AIC is not meaningfully lower than the best of SEM/SLM (typically a difference of at least 2 AIC units is considered meaningful), the added complexity is not justified and the more parsimonious model is preferred. Conversely, a clearly lower AIC for the SDM would suggest that neighbors' GDP directly matters for own life expectancy, beyond outcome diffusion alone.