Part A

Question 1: Which areas of Sweden have a significant (at the 5% level) spatial clustering of high or low values of income?

Map 1. LISA Spatial Autocorrelation Map of ln_income (QGIS, 2015)
This map displays the spatial clustering of income across municipalities using the Getis-Ord Gi* statistic, calculated in QGIS through the Visualist plugin. The Gi* statistic is presented as **z-values**, and the color gradient indicates where statistically significant clusters of high or low values occur.

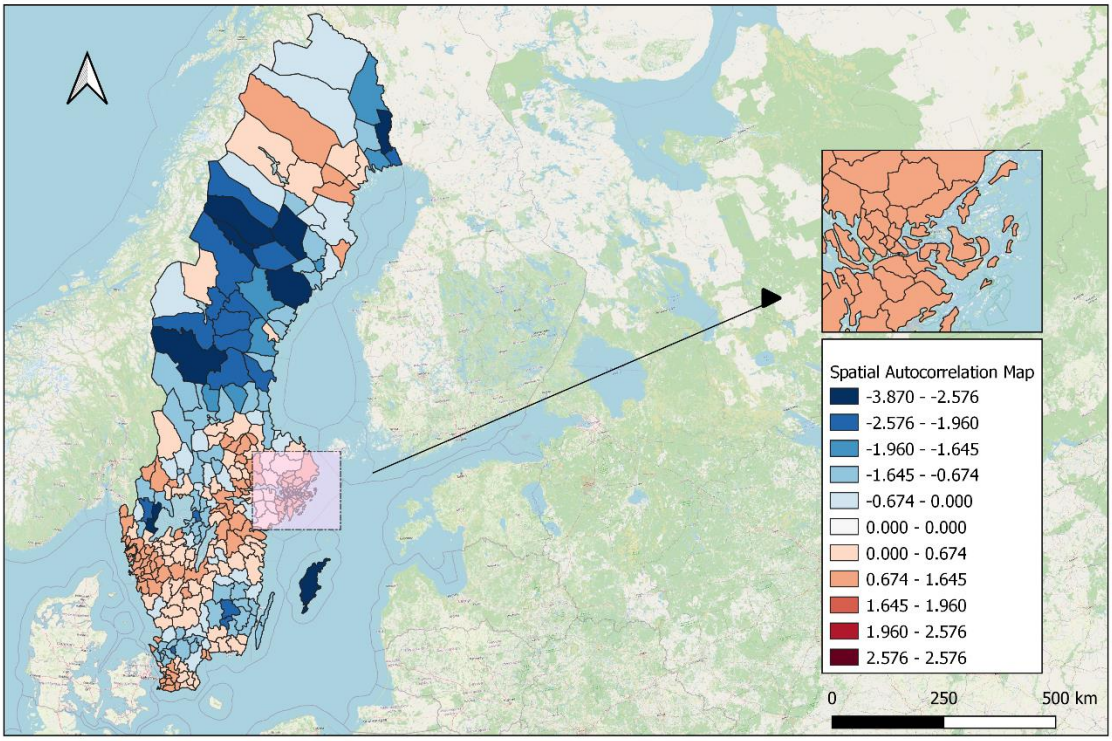
Interpretation of the z-values:

Red and Z ≥ 1.645: Statistically significant **positive clustering** of income (at the 10% level). These are areas where high-income values are surrounded by other high-income municipalities — so-called **hot spots**.
The **darker the red**, the **higher the z-score** and the **stronger the concentration** of high values.

Blue and Z ≤ –1.645: Statistically significant **negative clustering** of income (at the 10% level). These areas are **cold spots**, where low-income municipalities are surrounded by other low-income areas.
Darker blue means a **lower z-score** and a **stronger cluster of low values**.

White or pale areas: Not statistically significant (z-scores near 0), indicating no strong local clustering of income.

The **Stockholm region** is highlighted with a zoom-in frame, clearly showing a **hot spot** of high-income municipalities. This layout includes a legend, scale bar, north arrow and was constructed following lab instructions in QGIS.



“This LISA map visualizes how logged income is spatially clustered. Strong high-income clustering is observed around Stockholm, while low-income clustering appears in northern municipalities. The spatial pattern will be confirmed through global and local spatial statistics in Part B.

Part B

Question 2 Does the Global Moran’s I indicate significant spatial autocorrelation?

Null hypothesis (H_0): There is no spatial autocorrelation in the distribution of log average income across Swedish municipalities. The observed spatial pattern is random.

Alternative hypothesis (H_1): There is a statistically significant spatial autocorrelation in log average income — i.e., similar income values are geographically clustered.

Using GeoDa, we calculated the global Moran’s I for ln_income with the Queen contiguity weights matrix and got these results:

Moran’s I value: 0.546

Expected I (E[I]): -0.0035

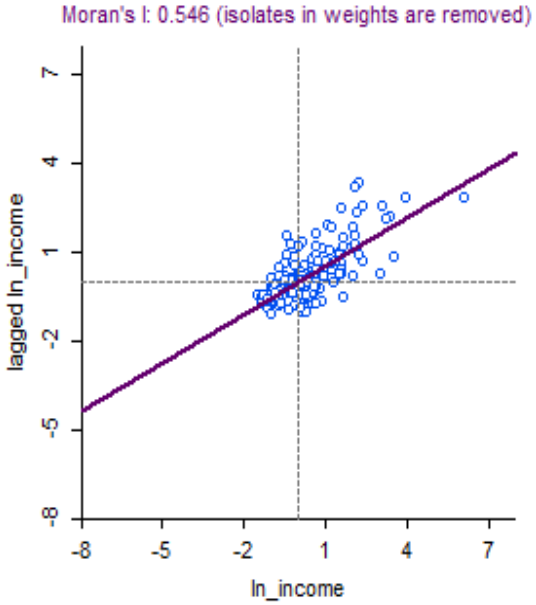
Standard deviation: 0.037

Z-value: 14.8438

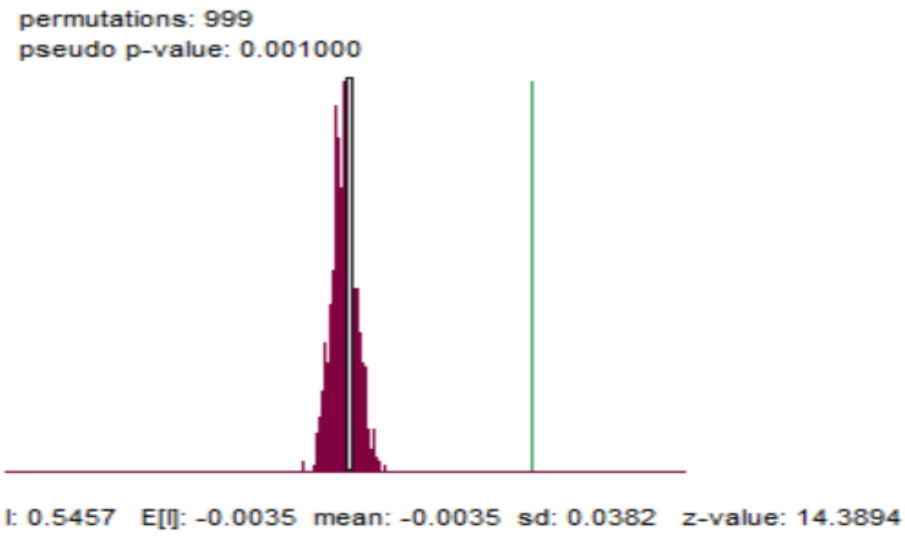
p-value (999 permutations): 0.001

At the 5% significance level, the p-value (0.001) is much smaller than 0.05, and the Z-value (14.84) lies far outside the critical region for a standard normal distribution.
So, we can reject the null hypothesis and conclude that there is strong evidence of positive spatial autocorrelation in municipal log average incomes in Sweden. Municipalities with similar income levels tend to be located near each other (either high-high or low-low)

Map 2. Moran's I scatterplot of ln_income (GeoDa).
This scatterplot shows a positive relationship between income and spatially lagged income, suggesting that municipalities with high income are often surrounded by other high-income ones. The Moran's I value is 0.546, which indicates strong positive spatial autocorrelation.



Map 3. Monte Carlo permutation test for Moran's I (GeoDa).
This histogram shows the results of 999 permutations used to test the significance of the Moran's I value. The observed value is far to the right of the distribution, and the pseudo p-value is 0.001, which confirms that the spatial autocorrelation is statistically significant.

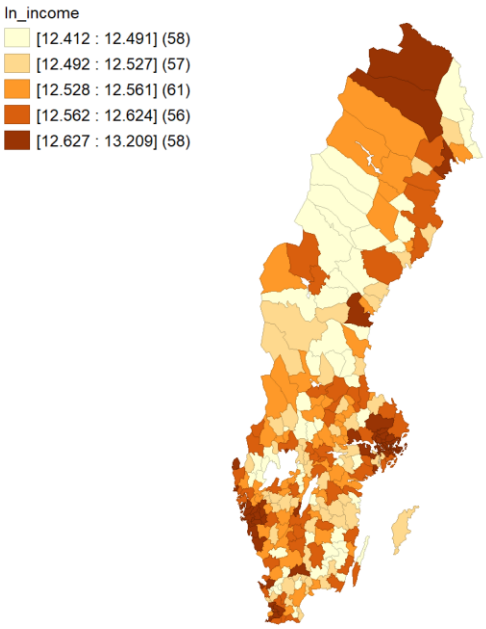


Question 3 – Where are the clusters located?

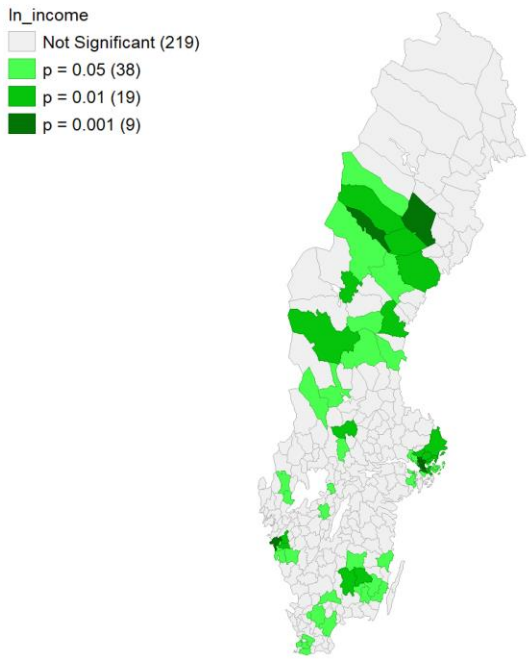
Map 4. Log of Average Income per Municipality (2015) (GeoDa) (Used as input for LISA analysis)

This map shows the raw log income values that were used as input in the LISA analysis displayed in Map 1

The darker areas have higher income than the lighter areas:



Map 5. LISA Significance Map of ln_income (GeoDa).
This map shows which municipalities have statistically significant local autocorrelation based on the p-values. Darker green areas are more significant, while grey areas are not significant.

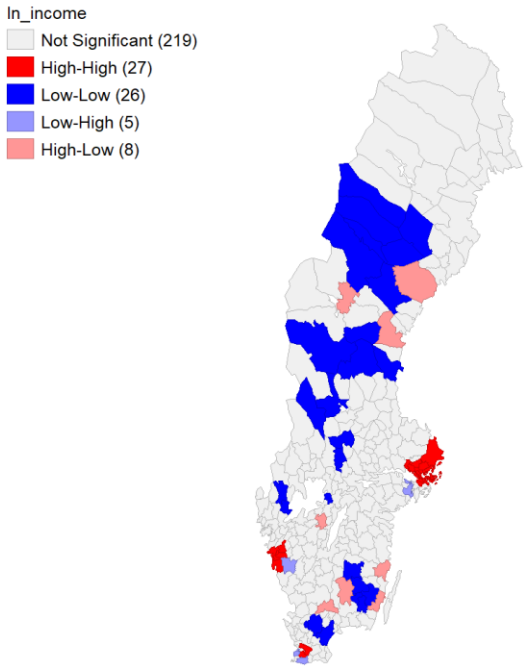


Map 6. LISA Cluster Map of ln_income (GeoDa).
This map identifies the type of clusters: High-High (red) and Low-Low (blue), where income levels are similar to neighbors. There are also outlier types like High-Low and Low-High, which are shown in pink and light blue. Most high-income clusters are located in and around Stockholm, and low-income clusters are more spread across central and northern Sweden.

High-High (Red): Stockholm and the surrounding areas make up a solid high-income group, showing how economic growth tends to gather in one place.

Low-Low (Blue): Northern and rural municipalities show persistent low-income clustering, possibly due to limited job opportunities or infrastructure.

Outliers (Pink/Light Blue): These areas (e.g., high-income municipalities bordering low-income ones) may indicate neighborhood transformation or regional disparities.



Part C:

1. OLS Regression (Classic Linear Model) (GeoDa)

I ran an OLS regression using ln_income as the dependent variable and Average_h as the independent variable. The model showed a very high **R-squared (0.9871)**, meaning that variation in household size explains most of the variation in income levels across municipalities. The coefficient for Average_h is positive (**≈ 3.03e-06**) and significant (**p = 0.00000**), so a higher average household size is associated with a higher logged income.

But, the diagnostic tests revealed some issues. The **Breusch-Pagan** and **Koenker-Bassett** tests both had p-values = 0.00000, indicating **heteroskedasticity** (non-constant variance of residuals). Also, the **Jarque-Bera** test had p = 0.00000, showing that the residuals are **not normally distributed**.

To test for spatial dependence, I ran LM and robust LM tests:

LM-lag = 12.8 (significant)
LM-error = 97.3 (highly significant)
Robust LM-lag = 101.6
Robust LM-error = 17.9

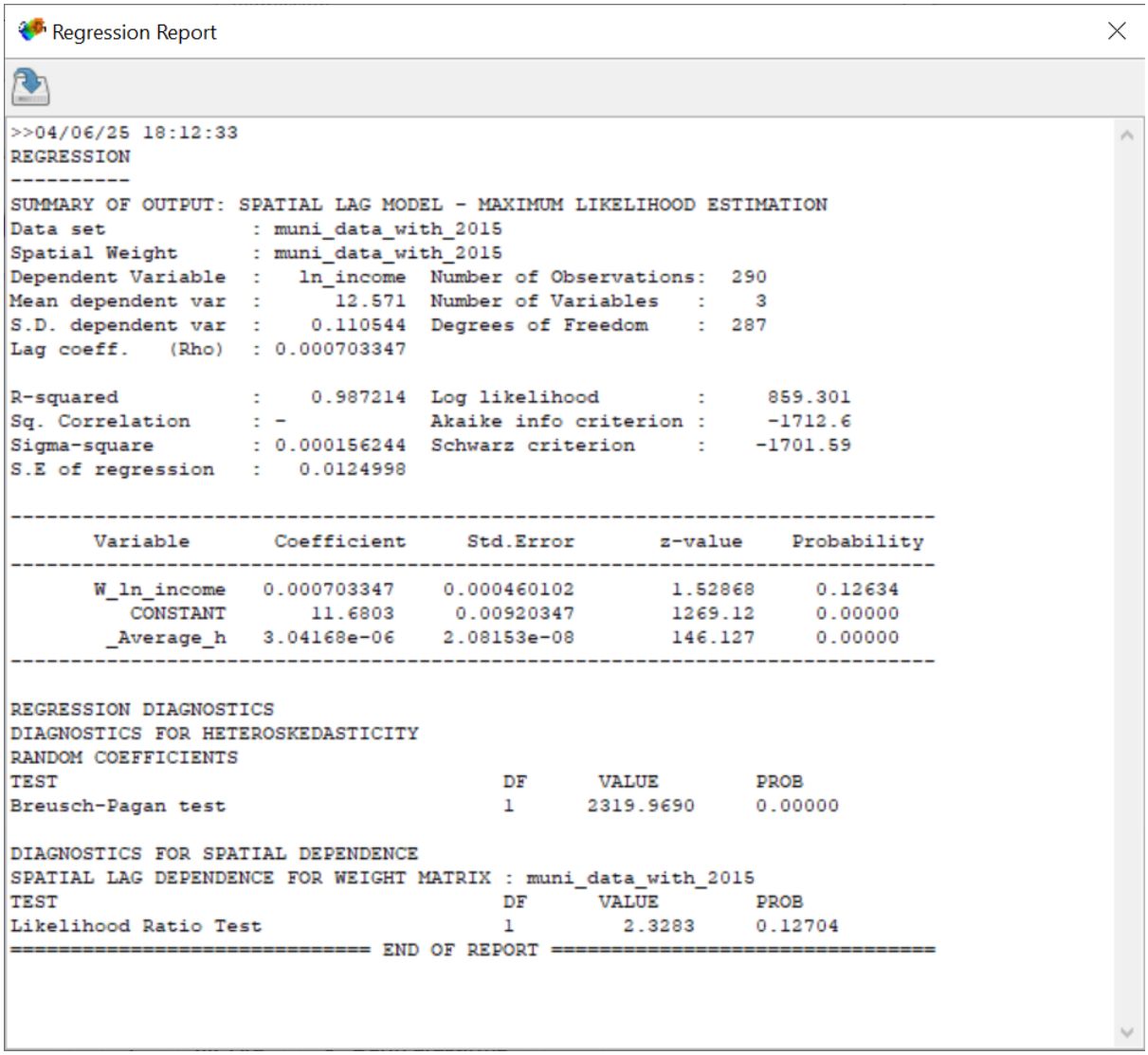
Since both robust tests are significant, but the **Robust LM-lag is much larger**, this suggests that the **Spatial Error Model (SEM)** is more appropriate than the lag model.

```
>>02/23/24 15:48:25
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  MuniData
Dependent Variable  :  _ln_Averag  Number of Observations:  290
Mean dependent var  :     12.9772  Number of Variables   :    2
S.D. dependent var  :    0.152321  Degrees of Freedom    :  288

R-squared           :    0.115676  F-statistic           :     37.6727
Adjusted R-squared  :    0.112606  Prob(F-statistic)     : 2.76359e-09
Sum squared residual:     5.95018  Log likelihood        :     152.044
Sigma-square        :   0.0206603  Akaike info criterion :    -300.089
S.E. of regression  :    0.143737  Schwarz criterion     :    -292.749
Sigma-square ML     :   0.0205179
S.E of regression ML:    0.143241

------------------------------------------------------------------------------
       Variable      Coefficient       Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
         CONSTANT       12.9623     0.00878209        1475.99     0.00000
        _pop_densi   0.000106961    1.74266e-05        6.13781     0.00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   1.327816
TEST ON NORMALITY OF ERRORS
TEST                    DF            VALUE              PROB
Jarque-Bera              2            492.1626           0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF            VALUE              PROB
Breusch-Pagan test       1             43.9556           0.00000
Koenker-Bassett test     1             11.7285           0.00062

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : MuniData
   (row-standardized weights)
TEST                            MI/DF         VALUE           PROB
Moran's I (error)              0.3800        10.1363         0.00000
Lagrange Multiplier (lag)           1        12.8962         0.00033
Robust LM (lag)                     1        17.1910         0.00003
Lagrange Multiplier (error)         1        97.3154         0.00000
Robust LM (error)                   1       101.6102         0.00000
Lagrange Multiplier (SARMA)         2       114.5064         0.00000
================================ END OF REPORT ================================
```

2. Spatial Lag Model (SLM)

In the SLM model, I included a spatially lagged income variable (W_ln_income) to test whether a municipality's income is influenced by neighboring areas. The lag coefficient was not significant ($p = 0.126$), meaning spatial dependence is not strong enough in this model.

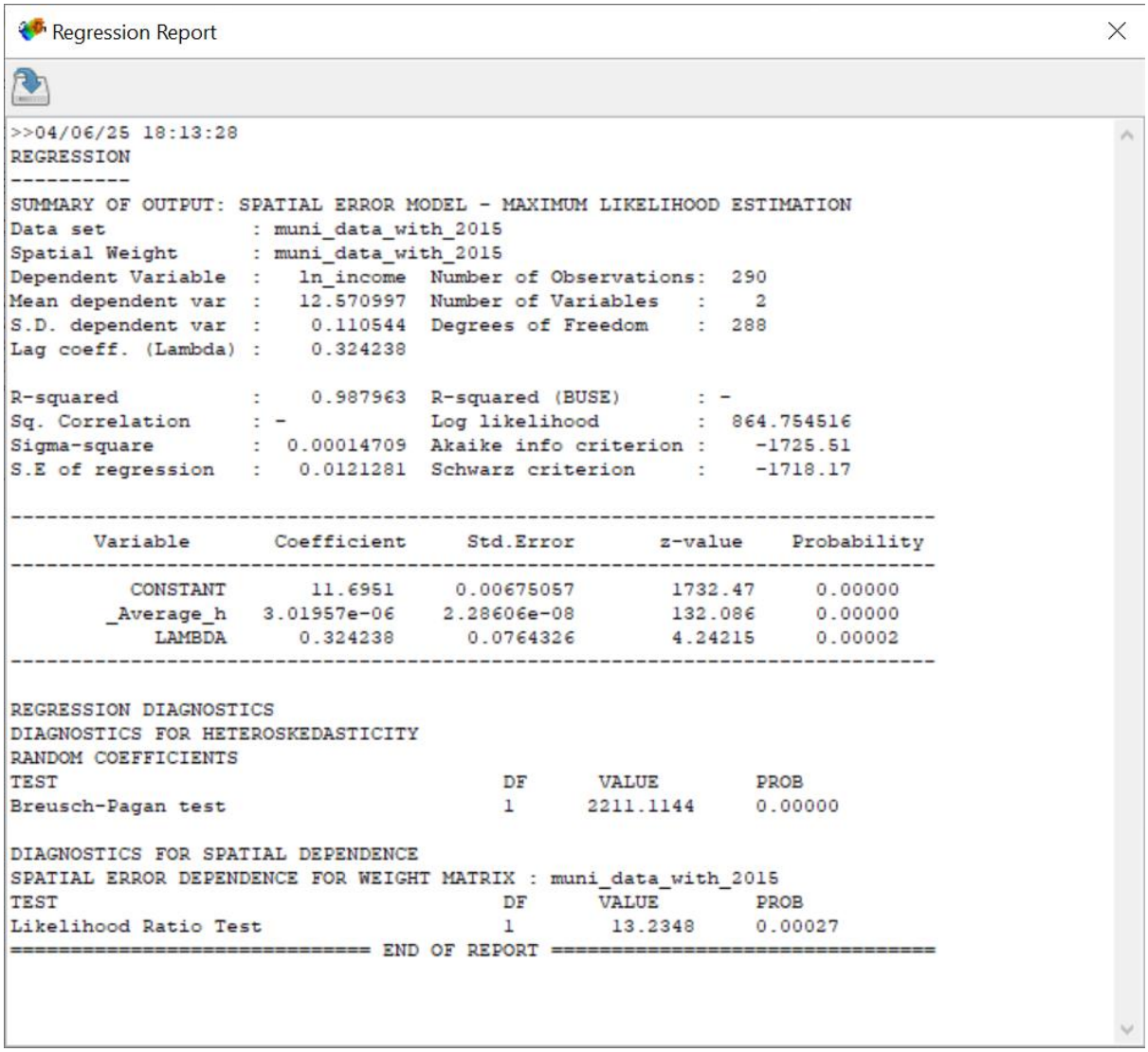
The model's $R^2 = 0.9872$, which is very close to OLS. But, heteroskedasticity remained an issue, and the model did not improve fit meaningfully compared to OLS.



3. Spatial Error Model (SEM)

The SEM addresses spatial autocorrelation in the error term, not the outcome variable. The spatial error coefficient $\lambda = 0.342$ was highly significant ($p = 0.00002$), confirming that OLS residuals had spatial dependence.

Additionally, the Likelihood Ratio test had $p = 0.00027$, showing improved model fit compared to OLS. Although heteroskedasticity still exists (Breusch-Pagan $p = 0.00000$), the spatial error model handles the spatial structure better than the SLM.



While OLS had a very high R^2, the residuals were not normally distributed and showed heteroskedasticity.
The spatial lag model didn't significantly improve the model, but the spatial error model corrected the spatial autocorrelation in the residuals.

Conclusion

Therefore, the **Spatial Error Model (SEM)** is the best option for modeling income in this dataset. The spatial error coefficient ($\lambda = 0.342$) is highly significant ($p = 0.00002$), which confirms spatial dependence in residuals. The robust LM tests showed better in SEM than in SLM (Robust LM-lag = 101.6 vs. LM-error = 17.9). And SEM handles heteroscedasticity better than OLS and SLM even though it didn't fully solve the non-normality of residuals. Future research could explore other covariates as for example: education, industry composition, etc.