**Spatial Statistics in**

**Middle East and Africa**

A05097101

Week 05

GSIAS, HUFS

Choi Doo Young Wicks

## Introduction and overview

In this lesson, we will explore the concept of **spatial dependence** and introduce two key regression models for analyzing it: the **Spatial Error Model (SEM)** and the **Spatial Lag Model**. These models are crucial tools in social science research, especially when addressing the interdependence between observations in spatial data.

Spatial dependence occurs when variables observed in neighboring or geographically close areas are related to each other. Failing to account for this correlation can result in inefficient or misleading results in statistical analysis. Thus, it is essential to properly incorporate spatial dependencies into your models when working with spatial data.

The SEM and the Spatial Lag Model handle spatial correlation differently:

* The **Spatial Error Model (SEM)** accounts for spatial correlation within the error terms.
* The **Spatial Lag Model** captures the impact of the dependent variable in one location on neighboring locations.

Understanding and applying these models is essential for producing accurate spatial analyses in the social sciences.

### The importance of spatial dependencies

Spatial dependencies exist in various fields, including **urban economics**, **demography**, and **environmental studies**. For instance, the economic performance of a particular region can be influenced by the economic conditions of neighbouring regions. These dependencies go beyond simple relationships between independent variables, reflecting more complex interactions that emerge in spatial patterns.

If these dependencies are not correctly modeled, estimates can become biased and inefficient. Specifically, suppose spatial correlation is ignored and Ordinary Least Squares (OLS) regression is used. In that case, the result may be poor model fit, and the variance of the residuals could be incorrectly estimated. Therefore, it is crucial to account for spatial correlation through models like the **Spatial Error Model (SEM)** or the **Spatial Lag Model** when analyzing spatial data.

These models differ in their approach:

* **SEM** addresses spatial correlation within the error terms.
* **The Spatial Lag Model** focuses on cases where the dependent variable is influenced by the values of the dependent variable in neighboring regions.

Choosing the appropriate model based on the data and the research question is essential to avoid biased or inefficient estimates.

#### Basic Concepts of the Spatial Error Model (SEM) and Spatial Lag Model

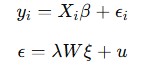
#### Spatial Error Model (SEM)

The **Spatial Error Model (SEM)** is used when spatial correlation exists in the error terms of the regression model. This model improves estimation efficiency by addressing the spatial correlation in the unexplained portion of the model, i.e., the residuals.

**Key concepts:**

* SEM captures **spatial correlation** in the error terms (residuals).
* When residuals from nearby observations exhibit similar patterns, it suggests the presence of spatial correlation.

The SEM can be expressed as follows:



Where:

* WW is the **spatial weight matrix**, which defines the spatial structure of the data (e.g., proximity or adjacency).
* λλ is the parameter representing the degree of spatial correlation.
* uu is the independent and normally distributed error term.

**Application:** The SEM is suitable when spatial correlation is found in the error terms, even if the explanatory variables are well-defined. It corrects for spatial dependencies that are not captured by the independent variables, thus improving model efficiency.

#### 2. Spatially Lag Model

The **Spatial Lag Model** is used when the values of the dependent variable in neighbouring regions directly influence the dependent variable. In other words, this model captures **spatial feedback**, where changes in one location affect neighbouring locations.

**Key concepts:**

* The **spatial dependence** of the dependent variable, meaning *yi* affects *yj*, where *j* represents a neighboring observation.

The model can be expressed as:



Where:

* ρis the **spatial autoregressive coefficient**, representing the influence of neighboring observations on the dependent variable.
* *Wy is the spatially lagged dependent variable, indicating how y in one region is affected by the values of y in neighbou*ring regions.

**Application:** The Spatial Lag Model is useful when the research question involves the transmission of a phenomenon across regions, such as housing prices or crime rates. For example, house prices in one area might be influenced by the prices in nearby areas.

### Differences Between the Spatial Error Model (SEM) and Spatial Lag Model

The **Spatial Error Model (SEM)** and the **Spatial Lag Model** are both used to handle spatial dependencies in data, but they approach the issue in distinct ways. Below are the key differences between these two models:

**1. Handling Spatial Dependencies:**

* **SEM:**
  + The SEM addresses **spatial correlation** in the **error terms**. This model is used when unexplained residuals are spatially correlated, rather than there being a direct interaction between dependent variables across regions.
  + For instance, when analyzing economic growth rates, if the unexplained variance (the error term) is spatially correlated between neighboring regions, SEM is used to correct for this spatial dependency.
* **Spatial Lag Model:**
  + The Spatial Lag Model captures the **direct interaction** between the dependent variable in one region and the dependent variable in neighboring regions. If the value of the dependent variable in one area affects the value of the dependent variable in a neighboring area, the Spatial Lag Model is more appropriate. This model explicitly models **spatial feedback** between dependent variables across regions.
  + For example, housing prices in one area could influence the housing prices in a neighboring area. The Spatial Lag Model captures this spatial spillover effect.

**2. Model Structure:**

* **SEM:**
  + The SEM models spatial correlation **only in the error terms**. It does not assume any direct spatial interaction between the dependent variables themselves.
  + Formula: y=Xβ+λWϵ+uy=Xβ+λWϵ+u
  + The focus is on capturing the spatial structure in the unexplained variance (error terms).
* **Spatial Lag Model:**
  + The Spatial Lag Model directly models the **spatial dependence** of the dependent variable itself. The model assumes that the value of the dependent variable in one region influences the value of the dependent variable in neighboring regions.
  + Formula: y=ρWy+Xβ+uy=ρWy+Xβ+u
  + This model emphasizes **spatial autoregression**, where changes in one location influence outcomes in nearby locations.

**3. Application Context:**

* **SEM:**
  + Suitable when the independent variables explain most of the variation in the dependent variable, but there is still spatial correlation in the error terms.
  + Example: Analyzing unemployment rates where well-defined independent variables are used, but neighboring regions have correlated residuals due to unobserved spatial factors.
* **Spatial Lag Model:**
  + Appropriate when the **dependent variable** in one region directly affects the **dependent variable** in neighboring regions.
  + Example: In a study on crime rates, where an increase in crime in one neighborhood could influence the crime rates in neighboring areas.

### Applications for SEM and Spatially Lag Models in Social Science Research

#### 1. Use cases for the Spatial Error Model (SEM)

The **Spatial Error Model (SEM)** is frequently applied in social science research when spatial correlation exists in the error terms, even if the independent variables are well-specified. By removing spatial correlation from the residuals, SEM improves model efficiency and produces more accurate estimates.

**Example:**

* **Unemployment and Economic Growth Analysis**
  + The explanatory variables (e.g., GDP, investment rates) might be well-defined when studying the relationship between economic growth and unemployment rates across different regions. However, neighbouring regions could still exhibit correlated error terms due to unobserved factors, such as regional labour market conditions or shared economic policies.
  + In this case, the SEM can be applied to adjust for the spatial correlation in the error terms and improve the overall model fit.

#### 2. Use Cases for the Spatial Lag Model

The **Spatial Lag Model** is used when the dependent variable in one location is affected by the dependent variable in neighbouring regions. This model is particularly useful for studying **spatial diffusion** or **local transmission** effects in crime rates, housing prices, or disease spread.

**Example:**

* **Crime Rates and Neighborhood Effects**
  + When studying crime rates, an increase in criminal activity in one neighborhood might lead to a spillover effect, increasing crime in neighboring areas as well. In this case, the Spatial Lag Model can be used to explicitly model the feedback effect between neighboring regions and measure the influence of spatial dependence.
  + The model allows us to capture how crime in one area influences crime in adjacent areas, revealing patterns of local transmission that would be missed by other models.

**Summary of Differences in Application:**

* **SEM:** Used when the main issue is spatial correlation in the error terms, without direct interaction between the dependent variables across regions.
* **Spatial Lag Model:** Applied when there is direct spatial feedback between the dependent variables, meaning that the values influence the values of the dependent variable in one region in neighboring regions.

These models can be effectively applied to various social science contexts, from urban studies to public health, where understanding the spatial relationships between regions is critical.

#### When to Use the Spatial Error Model (SEM) and How It Differs from the Spatial Lag Model

**1. How Spatial Dependencies Are Handled:**

* **SEM:**
  + SEM addresses spatial correlation by incorporating it into the **error term**. It corrects for situations where the **unexplained variance** (the error) is spatially correlated, but there is no spatial interaction between the dependent and independent variables.
  + Example: When analyzing economic growth, if the error terms are correlated between neighboring regions due to unobserved regional factors, SEM is used to remove this correlation.
* **Spatial Lag Model:**
  + The Spatial Lag Model, on the other hand, deals with **spatial interaction** between the dependent variables. It models feedback effects between the dependent variable in one region and the dependent variable in neighbouring regions.
  + Example: If house prices in one area directly affect house prices in a neighbouring area, the Spatial Lag Model captures this spatial feedback.

**2. Model Structure:**

* **SEM:**
  + SEM focuses on modelling spatial correlation only in **error terms**. The dependent variables are treated independently, but the spatial correlation in the residuals is addressed.
  + 폰트, 텍스트, 타이포그래피, 화이트이(가) 표시된 사진

    자동 생성된 설명
* **Spatial Lag Model:**
  + The Spatial Lag Model explicitly models the **spatial dependence** of the dependent variable itself, assuming that the value of the dependent variable in one region affects the value of the dependent variable in neighbouring regions.
  + 폰트, 타이포그래피, 텍스트, 서예이(가) 표시된 사진

    자동 생성된 설명
  + This model is characterised by the **spatial autoregressive coefficient** ρ, which captures the feedback effects between neighbouring regions.

**3. Application Context:**

* **SEM:**
  + SEM is most appropriate when the independent variables are sufficient to explain the dependent variable, but spatial correlation remains in the error terms. It is often used when the residuals are spatially clustered, but there is no spatial interaction between variables.
  + Example: When analyzing unemployment rates across regions, where well-defined explanatory variables are used, but neighboring regions exhibit similar patterns in their residuals due to shared regional factors.
* **Spatial Lag Model:**
  + The Spatial Lag Model is used when the dependent variables have direct spatial feedback. It is applied when the dependent variable in one region influences the dependent variable in neighbouring regions.
  + Example: Housing prices in one region influence prices in neighbouring regions.

## 2. **Theory Behind the Spatial Error Model (SEM)**

### The theory behind the Spatial Error Model

**The Spatial Error Model (SEM**) is essential for dealing with spatial correlation between error terms (residuals) in spatial data analysis. While ordinary regression analysis assumes that the error terms are independent, errors between neighbouring regions or observations are likely correlated in spatial data. In this case, ignoring spatial correlation can lead to biased estimates or underestimated standard errors, leading to incorrect statistical inferences.

SEM addresses the problems caused by spatial clustering or adjacency by

This spatial correlation is incorporated into the error term of the model. This increases the efficiency of the model and provides more reliable estimates by eliminating the influence of spatially correlated error terms on the analysis. SEM is a particularly useful model when the explanatory variables adequately explain the variability in the dependent variable, but the error terms are still spatially correlated.

This section will discuss the mathematical structure and basic principles of the SEM, explain how the model handles spatial correlation, and the analytical benefits it can provide.

#### Basic concepts and theoretical background of SEM

1. How spatial correlation affects modelling

**"Spatial autocorrelation"** refers to the phenomenon of similar characteristics between neighbouring regions or observations. For example, house prices in neighbouring areas may show similar trends, or economic performance may show similar patterns between neighbouring cities. If you ignore these spatial correlations and run a standard regression analysis, your estimates may be biased or inefficient.

In a typical regression analysis, we assume that each observation is independent.

However, this assumption is likely to be broken when dealing with spatial data. Neighbouring areas influence each other, and if you don't account for this spatial dependence in your model, the results of your regression analysis can be skewed.

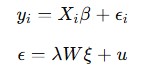
1. Issues and solutions when error terms contain spatial correlation

If the error terms include spatial correlation, it can lead to **underestimation** or **inefficient standard error estimation**. This occurs when error terms assumed to be independent because of spatial dependence are correlated. For example, if two neighbouring regions are affected by similar external factors and the model does not account for them, the error terms will be correlated.

To solve this problem, the **"Spatial Error Model (SEM**)" analyses spatial correlation by incorporating it into the error term. By handling spatial correlation in the error term, the SEM improves the model's efficiency and provides reliable estimates. The model explicitly models spatial correlation mathematically when it is present in the error term and removes its effects.

1. SEM's mathematical expression and the parameters' meaning (λ, etc.)

**The mathematical representation of the SEM** handles spatial dependence by including a parameter representing the error term's spatial correlation. The general formula for the SEM is



* + yi: Dependent variable
  + Xiβ: Independent variables and regression coefficients
  + i ϵ: error term
  + λ: parameter for spatial correlation
  + W: spatial weight matrix, indicating the adjacency between observations
  + ξ: Spatially correlated error
  + u: Independent error term (normal distribution)

The key to this model is a parameter called **λ**, which represents the degree of spatial correlation between the error terms. λ=0 means no spatial correlation, while λ≠0 ≠ the error terms are spatially correlated. W is the spatial adjacency matrix, which defines the spatial relationship based on the distance between two regions.

1. Difference between OLS and SEM: Problems when ignoring spatial autocorrelation

When using **"ordinary least squares" (OLS**), it is essential to assume that each observation is independent. However, in the presence of spatial correlation, this assumption is broken and can lead to a number of problems.

* 1. **Underestimation of standard errors**: Applying OLS while ignoring spatial autocorrelation leads to underestimation of standard errors because it does not account for correlation between error terms. This can lead to **misleading significance judgements**, i.e. variables that are not statistically significant may be incorrectly judged to be significant when in fact they are.
  2. **Efficiency issues**: OLS estimates can be inefficient if they ignore spatial autocorrelation. This means that estimates that account for spatial autocorrelation will have lower variance and produce more reliable results.
  3. **Potential for bias**: When spatial autocorrelation is present in the error term and ignored, OLS can provide biased estimates. In particular, if the error term reflects spatial factors that are not accounted for by the explanatory variables, the results may be skewed.

Therefore, it is important to use SEM to address these issues and obtain more accurate and reliable estimates in data with spatial autocorrelation.

### Comparison of Spatially Lag Model and SEM

#### Differences between the two models and where each applies

**The "Spatial Error Model (SEM**)" and the **Spatially Lag Model** are both spatial regression models that deal with spatial dependence, but they differ in the way they deal with this dependence and the circumstances in which they are applied.

* Spatial Error Model (SEM):

This model is used when the error terms are spatially correlated.

Explain why the errors in the model are spatially correlated, reflecting the spatial correlation included in the error term.

Situation: Used when independent variables are well defined, but unexplained external factors show spatially correlated patterns in neighbouring areas. For example, SEM is appropriate when unexplained economic changes are similar between neighbouring

cities.

**◦Key difference:** spatial correlation is reflected in the error term, and direct interactions between dependent variables are not considered.

* Spatially Lag Model:

◦Directly model spatial interactions of the dependent variables, i.e., a model that handles cases where the value of the dependent variable in a particular region affects the value of the dependent variable in neighbouring regions.

It is used when the dependent variable itself affects the value of the dependent variable in neighbouring regions through spatial interaction. For example, if house prices in one area are affected by house prices in neighbouring areas, a Spatially Lag Model is appropriate.

**◦ Key difference:** spatial autoregression (TERM) directly affects the dependent variable, thus reflecting spatial feedback between dependent variables.

#### Criteria for determining the suitability of each model

Criteria such as **log-likelihood**, **Akaike Information Criterion (AIC)**, and **R-squared** are commonly used to assess the goodness of fit of a model.

Log-likelihood **and** AIC are metrics that do not have **absolute** thresholds, but instead are interpreted through **comparisons between models**. This means that there is no absolute threshold above which these metrics are considered "good" or "bad". **Rsquared**, on the other hand, has a value between 0 and 1, and generally **closer to 1** indicates better explanatory power, so it can be considered to have a clear threshold. **log-likelihood:**

An indicator of how well the data fits a given model. The higher the value, the better the model can be interpreted as fitting the data.

By comparing the log-likelihood values of the two models, we can determine that the model with the larger value is a better fit to the data.

**∙Interpretation criteria**: The log-likelihood is interpreted as the higher the value, the better the model fits the data.

◦For example, **if** two models have log-likelihood values of -500 **and -480**, respectively, the model **with -480** is the better model.

◦The value itself shows the relative fit based on the comparison. **The larger the difference** between two models, the better the model is.

Akaike Information Criterion (AIC):

AIC is a metric that considers a model's goodness of fit and complexity. The lower the AIC value, the better the model is considered.

When comparing the AIC values between the SEM and the Spatially Lag Model, the model with the lower AIC value can be evaluated as a better fit.

**∙Interpretation criteria**: **The lower the** value of AIC, the better the model. This metric considers both the fit and complexity of the model.

◦For example, if two models have AIC values of 1200 and 1180, respectively, the model with 1180 is the better model.

If there is a difference of 10 or more between two models in AIC, the difference is considered significant, i.e., the larger the difference, the better the model.

R-squared:

An indicator of how well an independent variable explains the dependent variable, often used in general regression analysis. The closer the value is to 1, the higher the explanatory power.

In spatial models, R-squared is also useful for assessing the extent to which the variability in the dependent variable is explained. However, for spatial regression models, R-squared alone cannot fully determine the goodness of fit of the model and should be interpreted in conjunction with log-likelihood **or AIC**.

### Compare results when applying the two models with example data

**Example**: Compare the two models using a dataset that analyses the relationship between democracy and economic development. In this dataset, we are given the GDP and democracy level of each country, and we assume that proximity between countries plays an important role.

1. Apply Spatially Lag Model:
   * The dependent variable (level of democracy) is assumed to be influenced by the level of democracy in neighbouring countries.
   * In the model results, if the value of ρ **(spatial lag coefficient)** is significant, it means that the level of democracy in a country is spatially related to the level of democracy in its neighbouring countries.
   * **Interpretation of results**: If ρ is positive, it indicates that the higher the level of democracy in neighbouring countries, the higher the level of democracy in the country tends to be.
2. Apply Spatial Error Model:
   * Although the independent variable (GDP) fully explains the level of democracy, spatial autocorrelation may remain in the error term. Therefore, we apply SEM to address spatial autocorrelation in the error term.
   * **Interpretation of results**: If the value of λ (spatial error coefficient) is positive, it means that economic conditions between neighbouring countries impart spatial correlation to unexplained factors related to the level of democracy.
3. Compare results:
   * **Compare** log-likelihood **and AIC**: After fitting two models, compare the log-likelihood and AIC values to determine which model better fits the data. In general, a model with a lower AIC value and higher log-

likelihood is a better fit.

* + R-squared: By comparing the R-squared values of two models, you can see how well the independent variables explain the variability of the dependent variable. However, you should consider log-likelihood and AIC together rather than relying on R-squared alone to determine the goodness of fit of a model.

Summary

* **The Spatially Lag Model** is suitable for modelling direct spatial interactions between dependent variables, while the SEM is better suited to deal with spatial correlation between error terms.
* The goodness of fit between two models can be assessed through metrics such as **log-likelihood**, **AIC**, **R-squared**, and others, and when analysing data, these metrics should be considered collectively to select the most appropriate model.
* It is important to apply the two models in practice and analyse how each model produces different results.

## 3. lab session

### SEM analysis labs

◦ Explain how to implement the SEM model in R:

◦ Generate the spatial weight matrix (W) using the spdep package.

◦ Explain how to apply SEM to analyse and visualise results.

◦ Interpretation of results: meaning of λ values, determining whether error terms are spatially correlated, etc.

### Spatially Lag Model Lab

◦ Apply the Spatial Lag Model and compare the differences with SEM:

◦ Explain how to implement the model in R.

◦ Practice comparing and analysing the results of two models.

◦ Model evaluation: compare AIC, log-likelihood, etc. to determine which model is

the best fit.

Code description:

Apply both Spatial Error Model (SEM) and Spatially Lag Model (SAR).

* **SEM** is a model that takes into account spatial correlation in the error term.
* **SAR** is a model that assumes that the dependent variable in neighbouring countries affects the dependent variable in the current country.

1. Output and interpret model results:
   * Evaluate model fit by outputting **AIC** and **log-likelihood** values for **SEM** and **SAR** models.
   * Check the **lambda** (SEM) and **rho** (SAR) values to evaluate how the two models handle spatial correlation.
2. Compare models:
   * The model with the lower **AIC** value is considered to be the model that better explains the data, while the model with the higher **log-likelihood** value is evaluated as the better fit.
   * Provides a visualisation comparing the predictions of both models to actual GDP growth.
3. Visualisation:
   * You can visually compare the predictions of two models using **ggplot** and convert them into interactive graphs using **plotly**.

Interpret the full code and results

This code demonstrates how to analyse World Bank (WDI) data using the **spatial error model (SEM)** and **spatial lag model (SAR)**. This analysis evaluates and models **spatial correlations** between countries, with **GDP growth** as the dependent variable and **education expenditure** ratio **and population growth** as independent variables. The model analyses correlations but does not clarify causality.

Interpretation

1. preprocess data and create a spatial weight matrix
   * Download GDP, education spending, and population growth rates **from WDI data** to prepare spatial data by removing missing values and unrealistic coordinates.
   * Using each country's coordinates, generate **a spatial weight matrix** based on the K-Nearest Neighbours. This creates important underlying data that represents the spatial relationships between countries.
2. Apply a spatial error model (SEM)
   * SEM is a model that treats **spatial correlation** in the error term.
   * **The λ value** indicates the spatial correlation between the error terms. Evaluate the presence of spatial correlation by checking if the λ value is significant.

SEM model summary results:

* + **Lambda (λ)**: 0.29799, indicating that there is spatial correlation. **The p-value** is highly significant at 0.0049.
  + **The AIC value**: 1075.357 indicates the model’s goodness of fit, where a smaller AIC value means the model is a better fit.
  + **Log-likelihood value**: -532.6785 indicates how well the model explains the data.

1. Apply a spatial delay model (SAR)
   * The SAR model applies to the assumption that **the dependent variable** affects the dependent variable in neighbouring countries. You can evaluate this relationship through the **Rho** value.

SAR model summary results:

* + The **Rho value** is 0.29205, indicating that there is spatial correlation between the dependent variable (GDP growth) between neighbouring countries. **The pvalue** is 0.0048, which is highly significant.
  + **The AIC value** is 1075.327, indicating a slightly better fit model than SEM.
  + **Log-likelihood value**: -532.6637, indicating that the SAR model explains the data slightly better than the SEM model.

1. Compare the results of the two models:
   * **AIC**: The SAR model shows a lower AIC value (1075.327), making it the better fitting model.
   * Log-likelihood: The SAR model shows a higher log-likelihood value (-532.6637), indicating that it better explains the data.
2. Evaluate the model:
   * The SAR model is evaluated as the better-fitting model by AIC and loglikelihood. This suggests **that spatial interactions** are important, with **GDP growth rates** mutually influencing economic conditions in neighbouring countries.
3. visualisation:
   * It visually shows the relationship between **GDP** growth **rate, education expenditure ratio**, and **population growth rate** and compares the fit values of

the SEM and SAR models to give a visual indication of how each model explains the data.

Bottom line:

This code is useful for assessing correlation and modelling spatial correlation, but it cannot determine **causality**. When comparing the fit between the two models, the **SAR** model was rated as the better fit, but it focuses on explaining spatial correlation. It does not establish whether **education spending or population** growth is directly responsible for **GDP growth**. Establishing **causality** requires experimental design or more complex statistical methods.

## 4. Correlation is not causation

**The spatial error model (SEM**) and **spatial lag model (SAR)** are based on statistical **correlation** analysis. These models analyse **the** relationship **and spatial correlation between variables**, showing how certain variables are connected to other variables. However, these models do not directly prove **causality** on their own; however, they can provide the basis for inferring **causality** from the correlations.

Moving from correlation to causation:

1. **Correlation analysis**: This is the step where you check for interrelationships between two variables. This helps you understand the connections between them. For example, you might find a correlation between education spending and GDP growth, but this doesn't mean **that education spending caused GDP growth.**
2. **Form a causal hypothesis**: You can formulate **a hypothesis** based on the results

of your correlation analysis. For example, you might hypothesise that an increase in education spending will have a positive effect on GDP growth.

1. **Causal analysis methods**: To validate this hypothesis, we need to use additional **causal analysis** methods. We can use the following methods to do this
   * **Analyse panel data**: Track changes in data over time and test for causality through fixed- or random-effects models.
   * **Instrumental variable method (IV)**: To address the endogeneity problem, instrumental variables that provide exogenous shocks can be used to clarify causality.
   * **Difference-in-difference (DD) analysis**: You can analyse causality by comparing the effects before and after a policy change or event.
   * **Randomised experiment or quasi-experiment**: Experimental designs allow you to clarify the causal relationship between variables.
2. **Interpret causal analysis results**: Causal analysis allows you to clearly measure the impact of one variable (such as education spending) on another (such as GDP growth). This allows you to move beyond simple correlations and clearly explain **which cause led to which effect.**

Bottom line:

The correlation analyses you've performed so far are useful for identifying relationships between variables. They provide important foundational data that allows you to enter **causal analysis**. Causal analysis involves moving **from correlation to causation**, which requires further application of various statistical methodologies.