GEOG0114: PRINCIPLES OF SPATIAL ANALYSIS

WEEK 9: SPATIAL MODELS (PART 1)

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

Contents

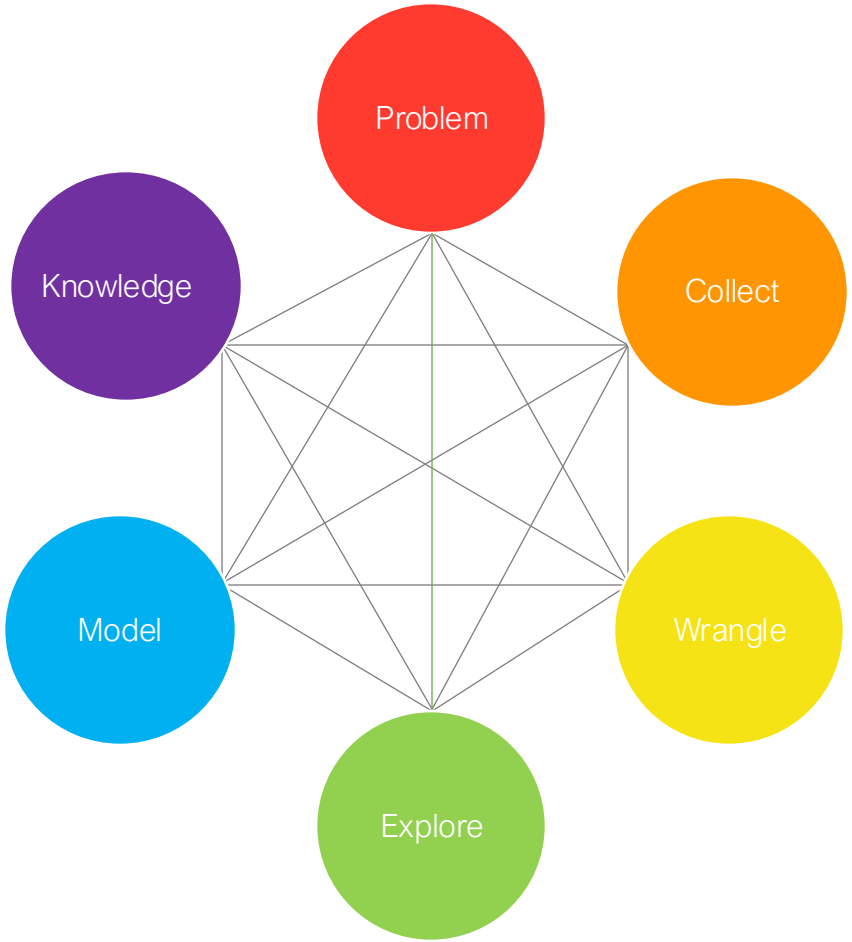
1. Regression-based models for causal and predictive inference

2. Standard Linear Regression (SLR)
   - Model Diagnostics of Residuals
   - Testing evidence for Spatial Dependence

3. Extending an SLR to Spatial Regression Model
   - Spatial lag
   - Spatial error

4. Methodology for model selection

5. We'll cover an example of analysis and interpretation



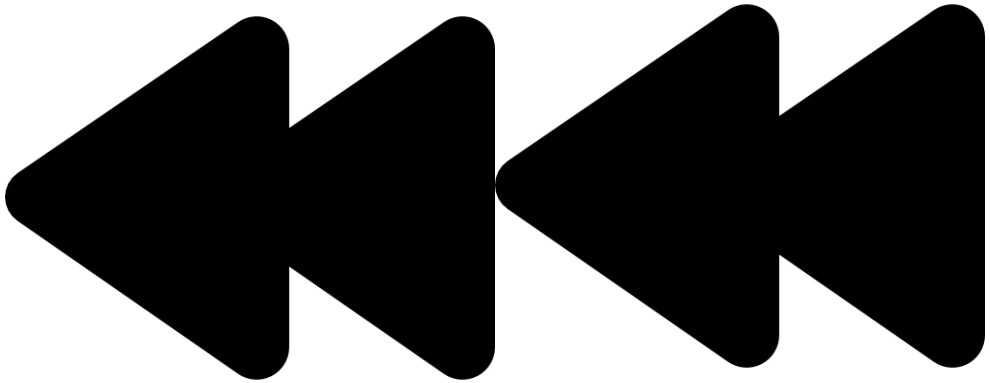
QUICK RECAP

1. We covered key concepts of spatial dependence i.e., the characteristics such spatial heterogeneity and autocorrelation

2. We covered what, why and how Spatial Weight Matrix (W) are important
   - Contiguity-based weights
   - Distance-based weights

3. How to measure globally and locally the degree of spatial dependence using the Moran's I statistics

Let's rewind a bit back on what was taught in Week 3



Spatial Models

Week 9

Spatial Lag and Error Models

Week 10

Geographically Weighted Regression (GWR) Models

Week 3's concepts on spatial dependence are heavily emphasised here for Spatial Lag and Error Models

What is a Regression model

Definition of a Regression model:

Regression is a mathematical formula that allows a user to perform two things:

1) To determine the **relationship** or **association** between a specified **outcome** (i.e., **dependent variable**) with one or more **predictors** (i.e., **independent variable(s)**)

2) For making a prediction about an outcome (i.e., **dependent variable**), or making a projection or forecast of that outcome based on predictors.

1) Casual Inference

2) Predictive Inference

Casual & Predictive Inference

Examples of causal inference, i.e., the relationship between a dependent and independent variable(s):

- What is the **association** between the incidence of skin cancer and exposure to long-term low-level elevated topsoil arsenic concentrations in the UK?
- What **impact** does increased levels of socioeconomic deprivation have on the burden of crime rates?
- What is the risk of a landslide **associated** with an increase in levels of rainfall, as well as soil texture in Bangladesh?

Examples of predictive inference, i.e., when we know relationship between dependent and independent variable(s) but use some model to forecast unobserved values:

- What is the predicted incidence rate of skin cancer in the UK (at residential post code-level) when topsoil arsenic concentrations reach levels of 40 mg/kg
- What is the estimated risk of a landslide event in Bangladesh when rainfall levels reach values > 50mm (2.0 inches) per hour

Let us clear the air with some terminology

Dependent Variable

This is the variable whose values we want to explain or forecast

Its value depend on something else

Interchangeable terms:

- **Dependent variable**
- **Response variable**
- **Outcome**
- **Event**

Independent Variable

This is the variable that explains our dependent variable. We can say it has an “impact”, or it is “associated” with our dependent variable

Its values are independent and used to explain the dependent variable

Interchangeable terms:

- **Independent variable**
- **Explanatory variable**
- **Exposure**
- **Risk factor(s)**

In terms of regression, there are several types of models, each with there own families depending on the type distribution for the dependent variable:

Here is a board overview:

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc.,	**Linear regression**
Binary measures (1 = "present" or 0 = "absent"): e.g., A person voting for a candidate, Lung cancer risk, house infested with mosquitoes etc.,	**Logistic Regression**
Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with mosquitoes, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc.,	**Logistic Regression**
Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc.,	**Poisson Regression**
Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc.,	**Survival Analysis with Cox regression**

In terms of regression, there are several types of models, each with there own families depending on the type distribution for the dependent variable:

Here is a board overview:

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc.,	**Linear regression**
Binary measures (1 = "present" or 0 = "absent"): e.g., Person's voting for a candidate, Lung cancer risk, house infested with rodents etc.,	**Logistic Regression**
Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with rodents, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc.,	**Logistic Regression**
Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc.,	**Poisson Regression**
Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc.,	**Survival Analysis with Cox regression**

Univariable & Multivariable Regression

A Linear Equation

- Do you remember one of these?
  - $\boldsymbol{y = a + bx}$
  - $\boldsymbol{y = mx + b}$

- In the statistics world, we just use a different notation:
  - $\boldsymbol{y = \beta_0 + \beta_1 x}$

Components of the above linear equation:

y is the dependent variable
x is the independent variable
$\boldsymbol{\beta_0}$ is the intercept of the line on the y-axis when ***x*** is 0
$\boldsymbol{\beta_1}$ is the slope of the line. It is amount of how ***y*** increases when there's a unit increase in ***x***

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Variables

- y is the dependent variable
- x is the independent variable

Parameters

- β_0 is the intercept
- β_1 is the slope, also called a coefficient
- ε is the error term

Multivariable Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

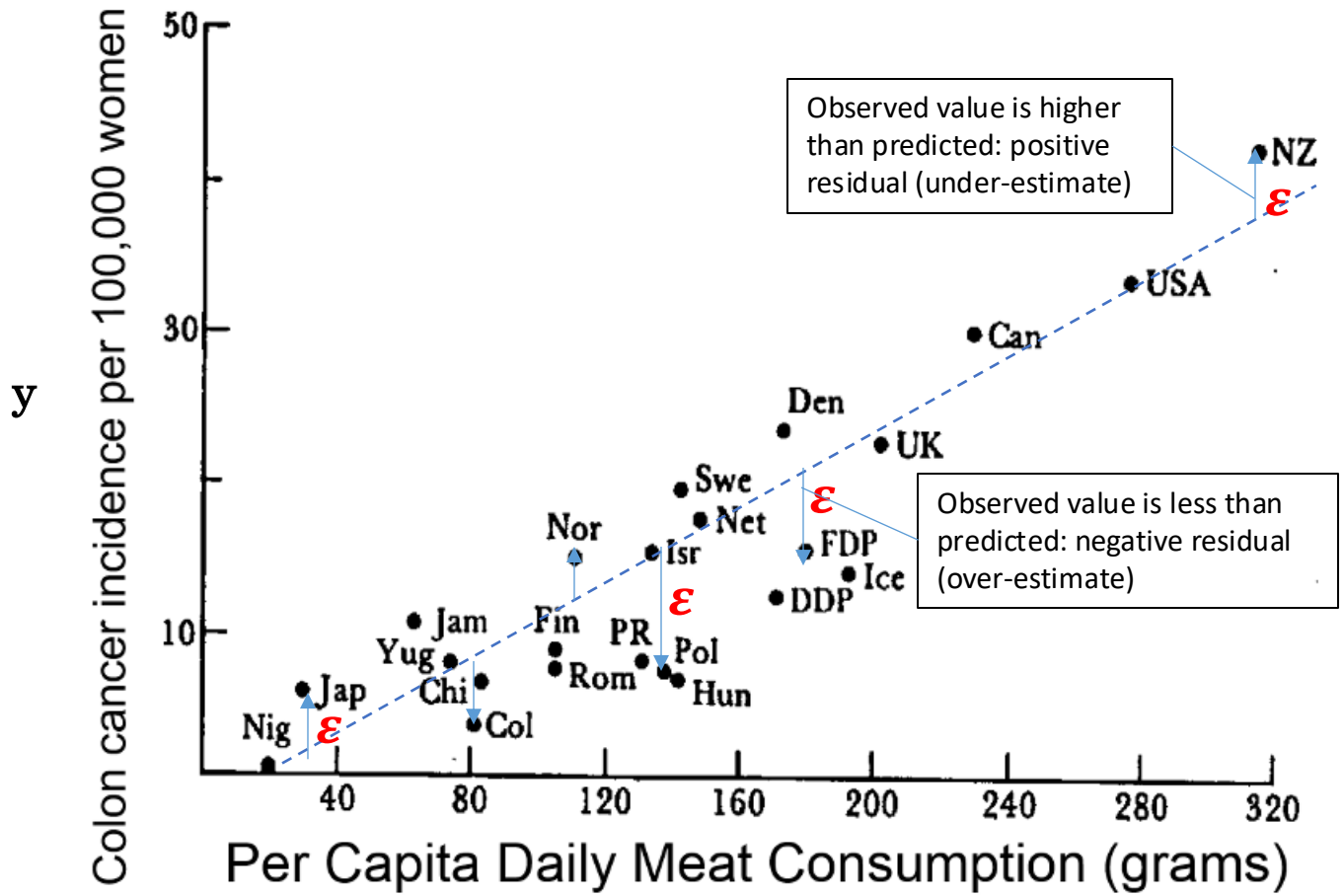
Variables

- y is the dependent variable
- $x_1, x_2, x_3, \ldots, x_k$ are the independent variables

Parameters

- β_0 is the intercept
- $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are the slopes (or coefficients) for the corresponding variables $x_1, x_2, x_3, \ldots, x_k$
- ε is the error term

Simplest example of a univariable relationship



1. Notes:

Linear models are best for finding the best fit between predicted values and observed values y

2. Notes:

a. Blue line – the predicted value for y and the blue hollow circles are the actual observed.

b. The difference between the predicted and observed value give us the error term which we term as residuals.

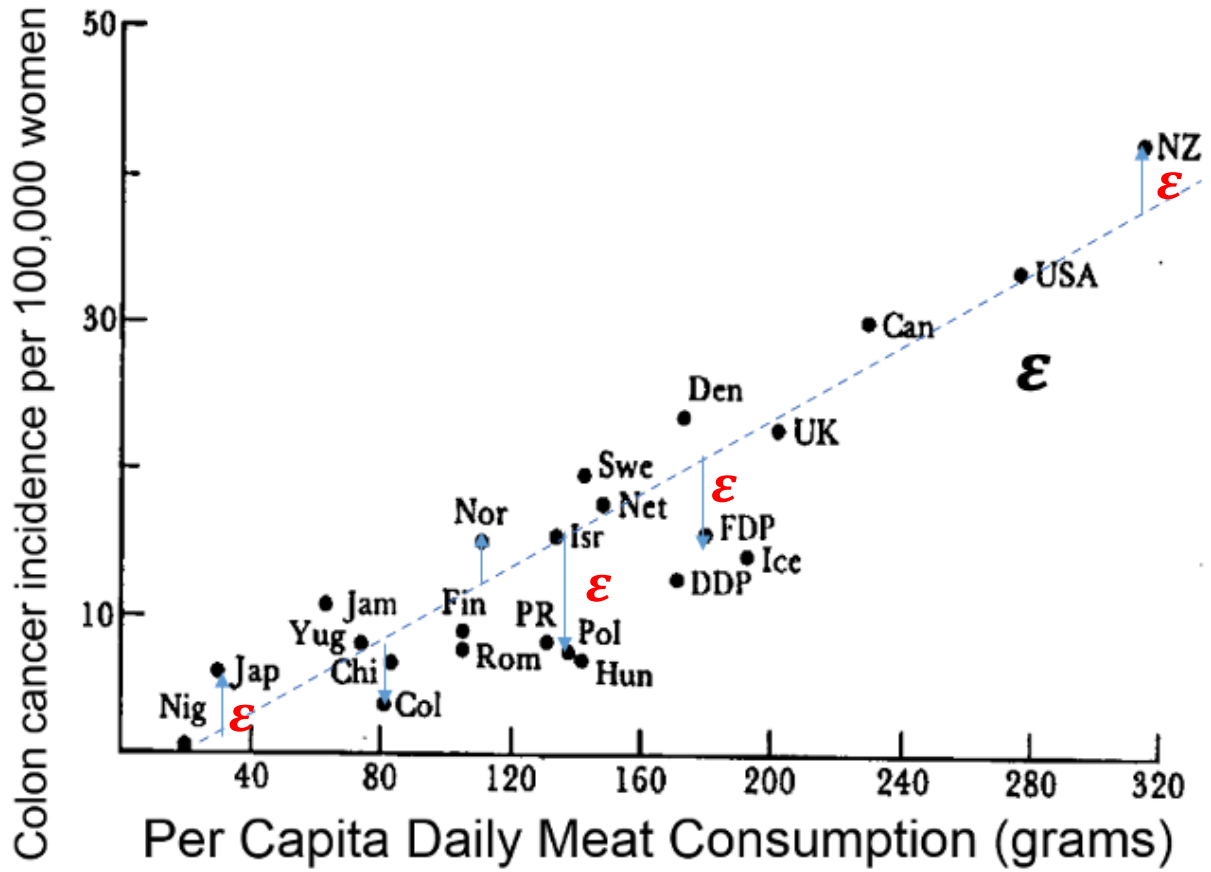
Such model comes with a set of assumptions

This model returns the actual value for y (Black dot): $y = \beta_0 + \beta_1 x_1 + \varepsilon$

This returns the fitted line (blue) which is a prediction for y: $\hat{y} = \beta_0 + \beta_1 x_1$

Assumptions of Linear Model

y



This model returns the actual value for y (Black dot): $y = \beta_0 + \beta_1 x_1 + \varepsilon$

This returns the fitted line (blue) which is a prediction for y: $\hat{y} = \beta_0 + \beta_1 x_1$

Notes:

<u>Assumption 1: The random errors/ residuals should have a mean of zero.</u>

Meaning that if we were to take the sum of these differences between the predicted y and observed y, and then divide it be the number of data points – this should give a mean estimate that should be zero.

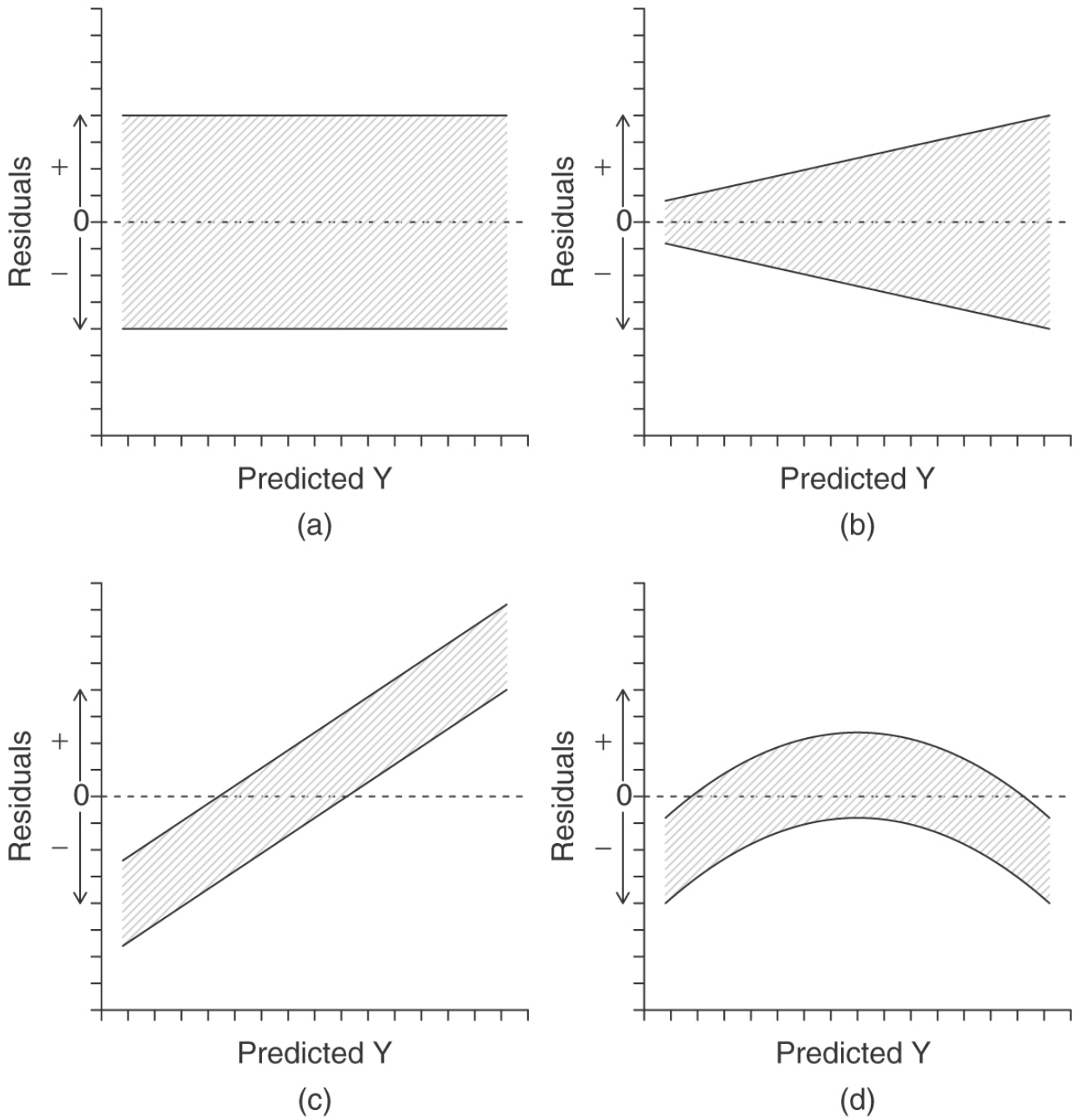
Well, this assumption is saying that this sum should be approximately close to zero.

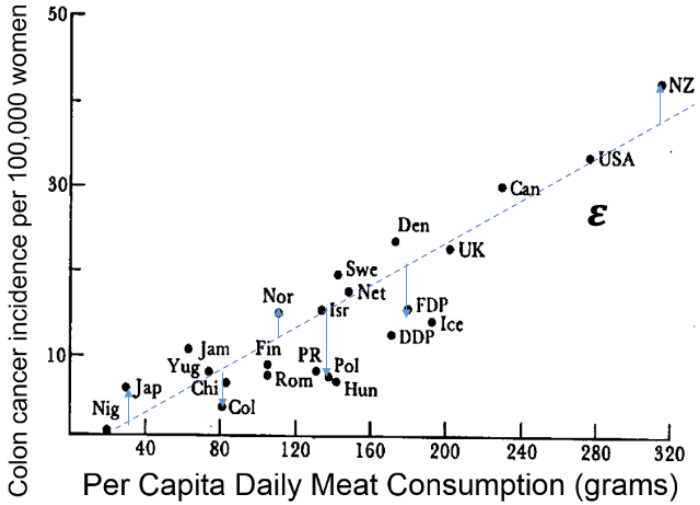
<u>Assumption 2: The distribution of these errors/residuals must be normally distributed</u>

Meaning that if we were to take these points and generate a histogram with them – these should give a shape that's approximately akin to a bell-curve.

<u>Assumption 3: The errors/residuals must have a constant variance (homoskedasticity) and must be uncorrelated.</u>

What do we mean by this third assumption???



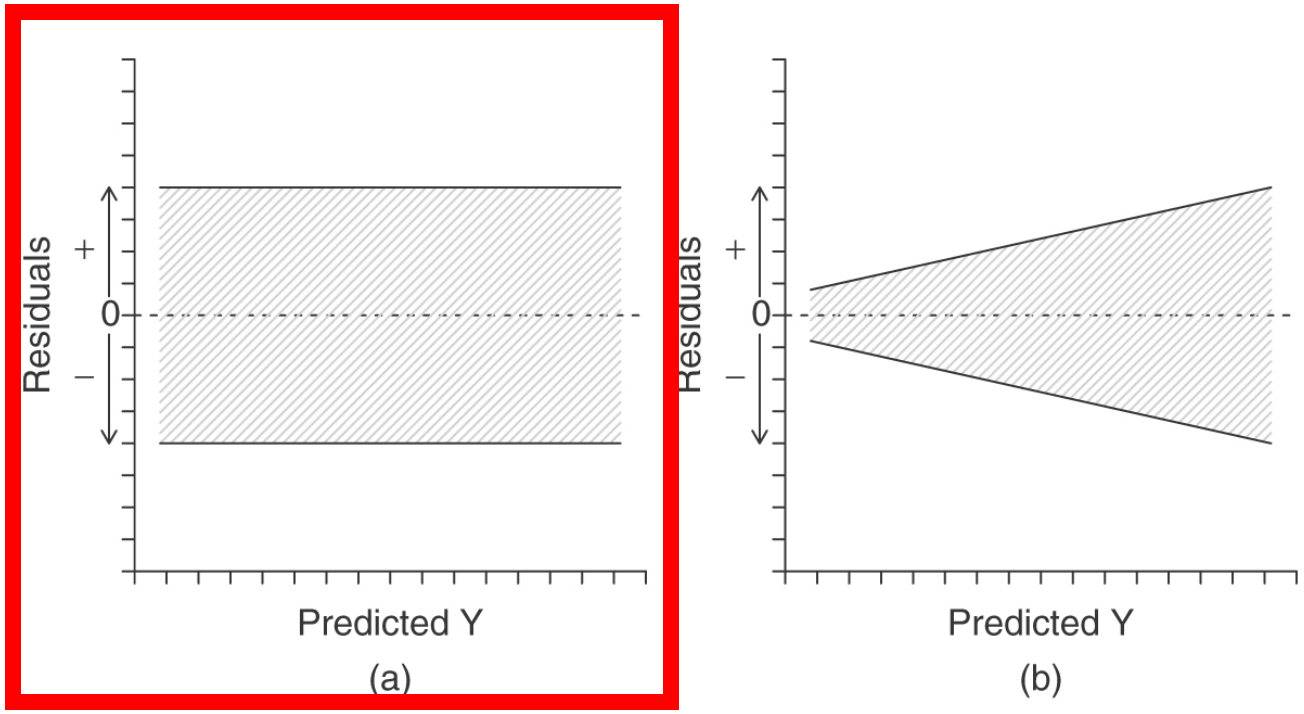


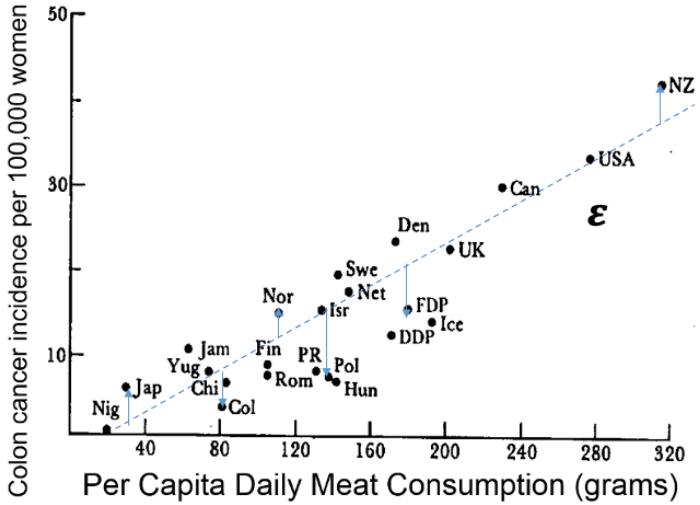
Notes:

<u>Assumption 3: The errors/residuals must have a constant variance (homoskedasticity) and uncorrelated.</u>

What do we mean by this assumption???

When we take the residuals and plot them against the predicted Y (i.e., values of the line of goodness of fit) they should exhibit a patterns that appear like the image in panel (a).





Notes:

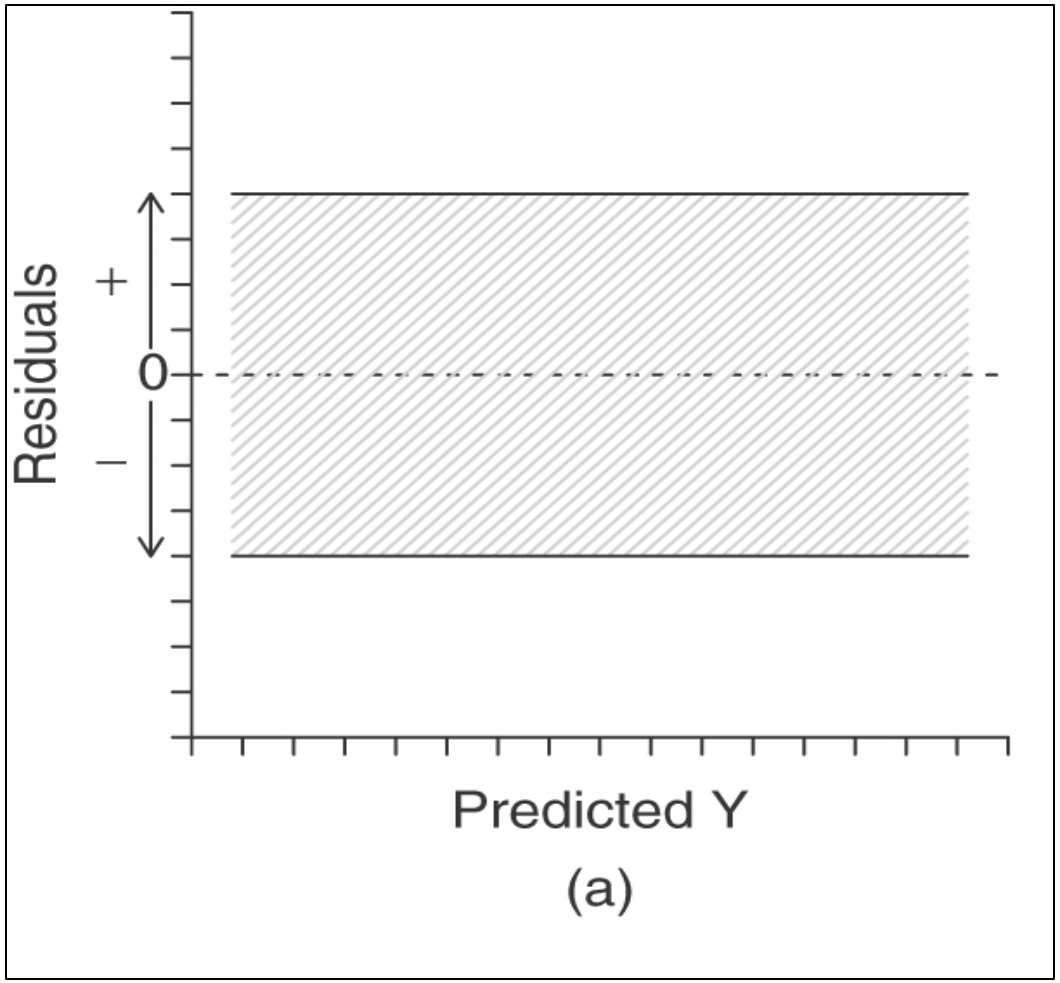
<u>Assumption 3: The errors/residuals must have a constant variance (homoskedasticity) and uncorrelated.</u>

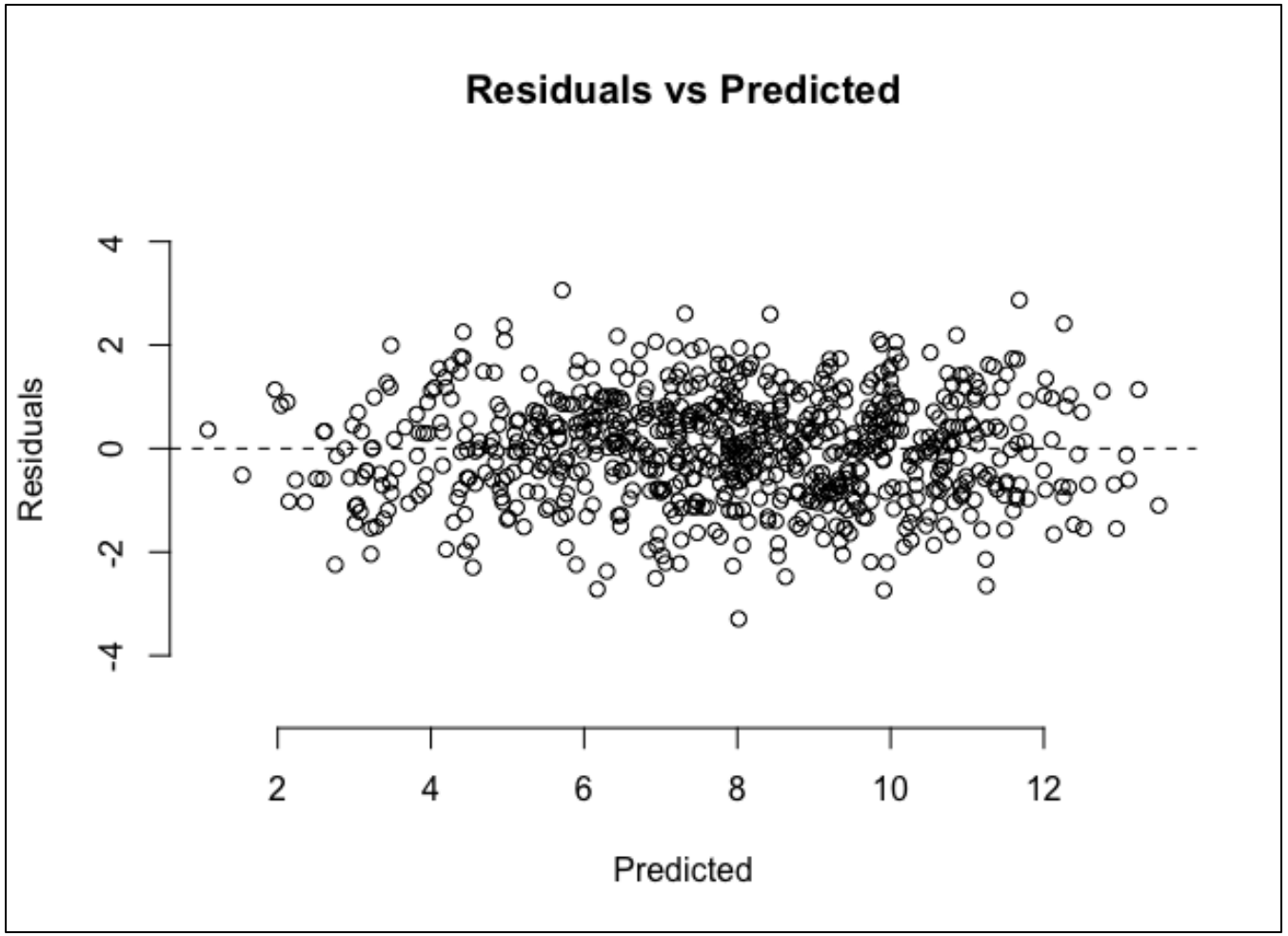
What do we mean by this assumption???

When we take the residuals and plot them against the predicted Y (i.e., values of the line of goodness of fit) they should exhibit a patterns that appear like the image in panel (a).

If the patterns appear like this, then it's safe to say that the variance in the errors are constant across the predicted values.

However, if is noisy like in panels b, c and d then there's evidence of <u>heteroskedasticity</u>



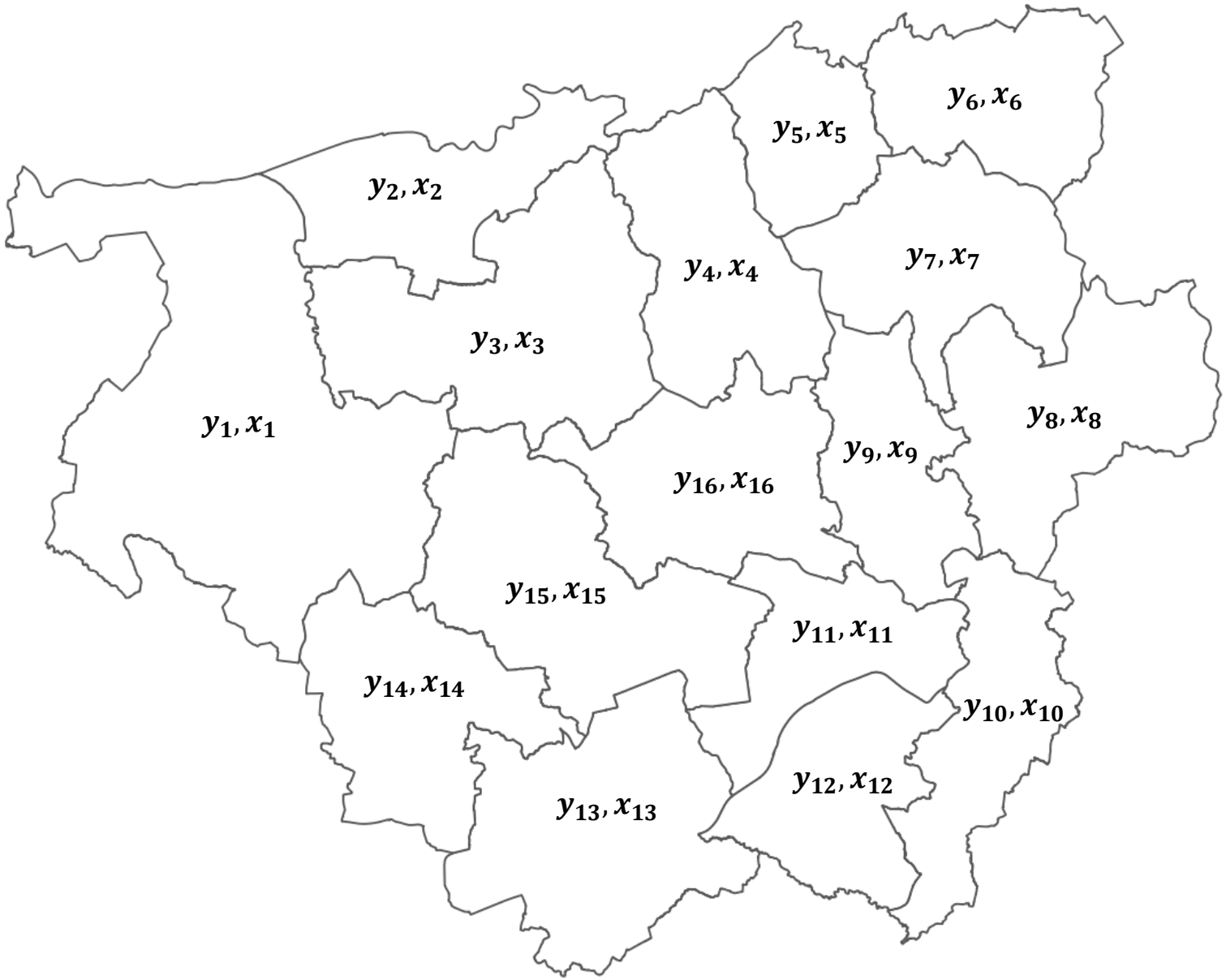


The points should be concentrated around the line when **y = 0**

Example plot of when the 3rd assumption is NOT violated

Spatial (Lag & Error) Linear Regression Model

Suppose we have a hypothetical study area with 16 areas



Notes:

Let Y be some dependent variable that is continuous and normally distributed, where there are 16 observation for Y (i.e. $\boldsymbol{y_1, y_2 \ldots y_{16}}$)

- For example: Averaged house price (£)

Let X be some independent variable, where there are 16 observations for X (i.e. $\boldsymbol{x_1, x_2 \ldots x_{16}}$)

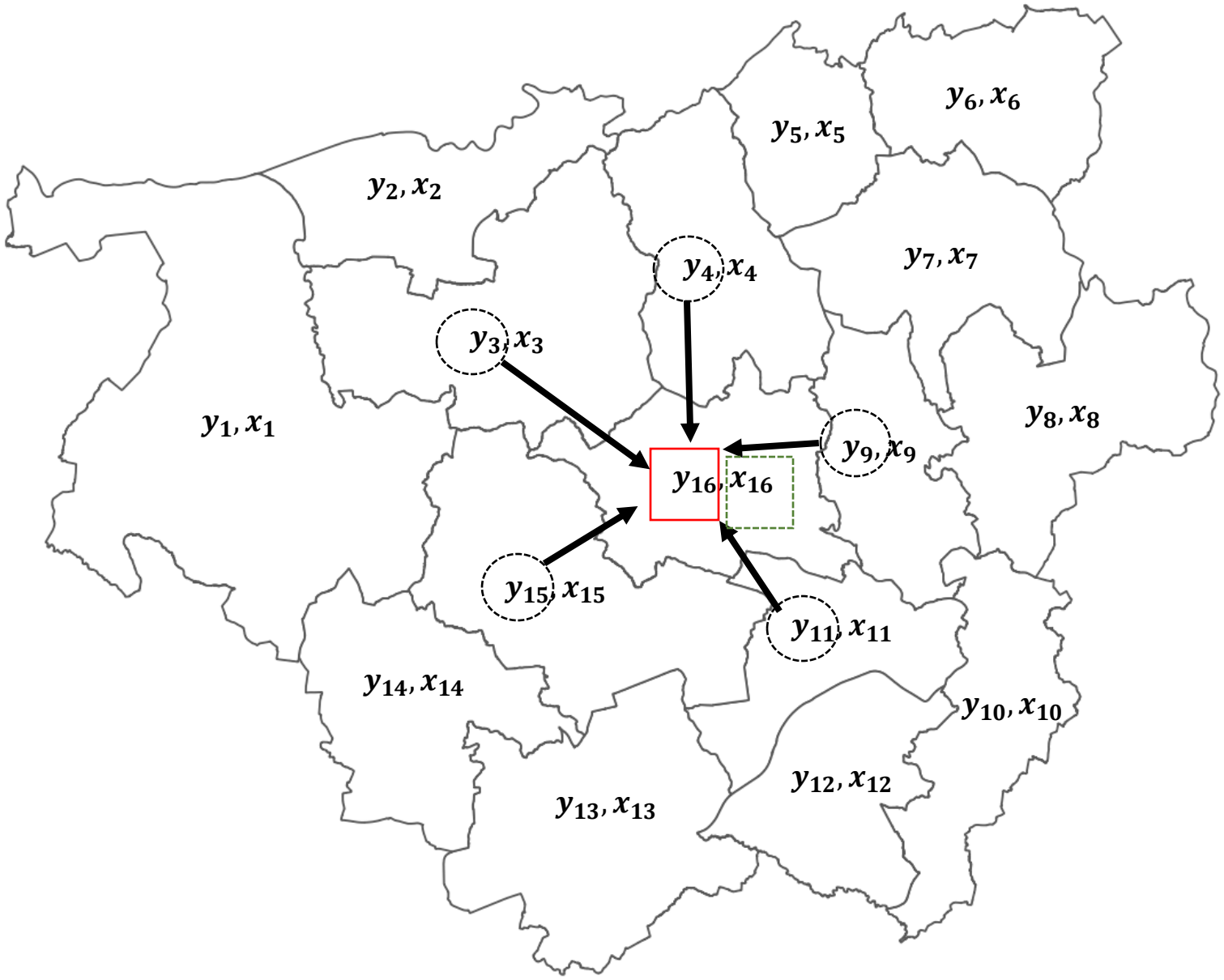
- For example: Socioeconomic deprivation score

Research question: To investigate the **geospatial** impacts of socioeconomic deprivation on house price in this hypothetical study area.

Insufficient to use the typical linear regression model for this context

$$\boldsymbol{y = \beta_0 + \beta x + \varepsilon}$$

Scenario 1: The outcome of a location is influenced by other outcomes values from direct neighbours [1]



A hypothetical study area with 16 areas

Notes:

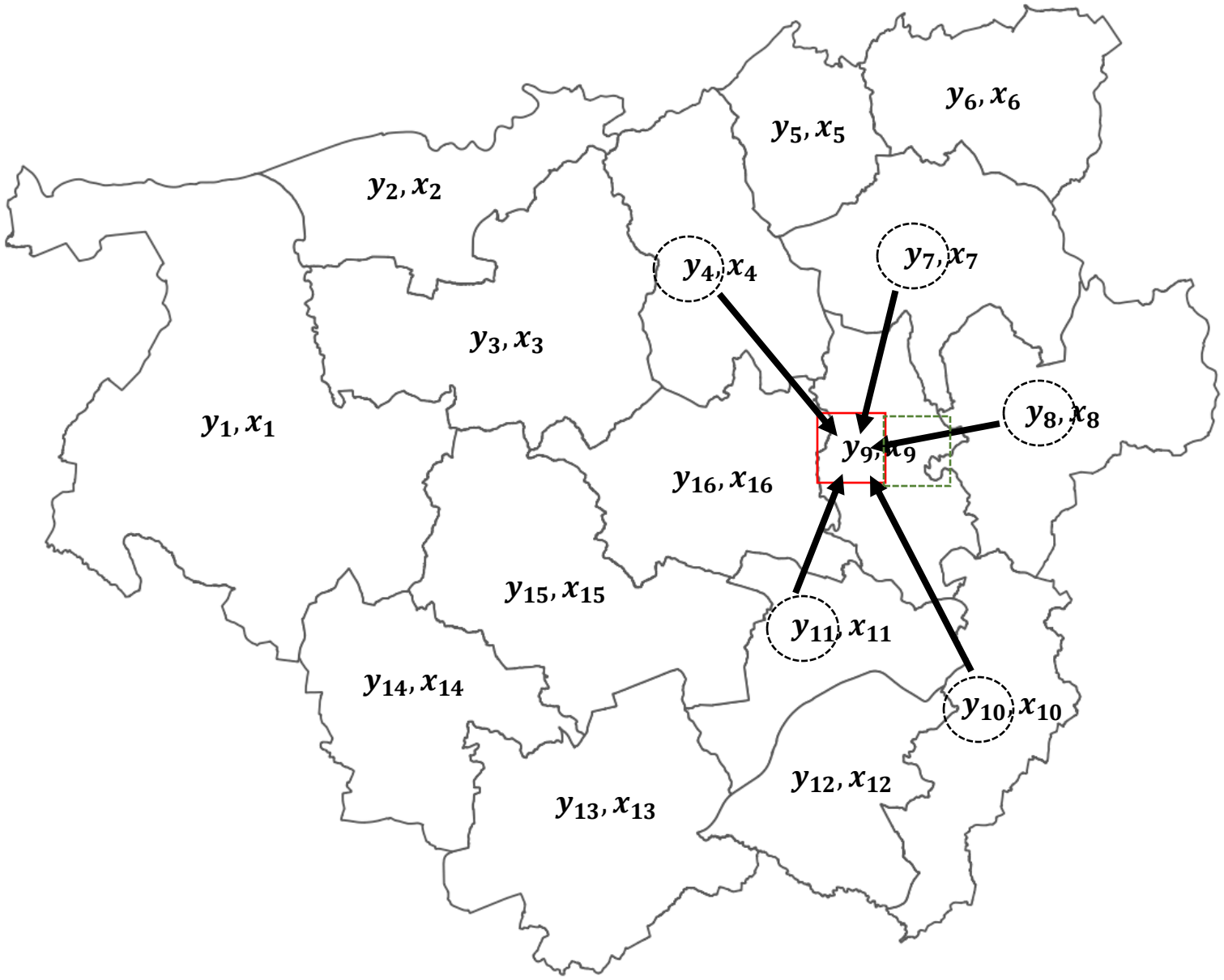
One type of spatial dependency in this neighbourhood structure can be seen when a **dependent variable (house price)** for some neighbourhood (16) (y_{16}) is influenced by other house price **dependent values** from neighbouring areas (i.e., house prices values y_3, y_4 , y_9, y_{11} and y_{15}) such as neighbourhoods 3, 4, 9, 11, and 16.

In this scenario, when the dependent variable's outcomes from neighbouring areas have an influence on the outcome of a location, we call this type of spatial dependence a **Spatial Lag on the dependent variable (aka Spatially lagged Y model)**

Spatial Lag are suggestive of spatial spill overs or diffusion.

With spatial lags in a linear regression, the assumptions about independence, and observations being uncorrelated are always violated.

Scenario 1: The outcome of a location is influenced by other outcomes values from direct neighbours [2]



A hypothetical study area with 16 areas

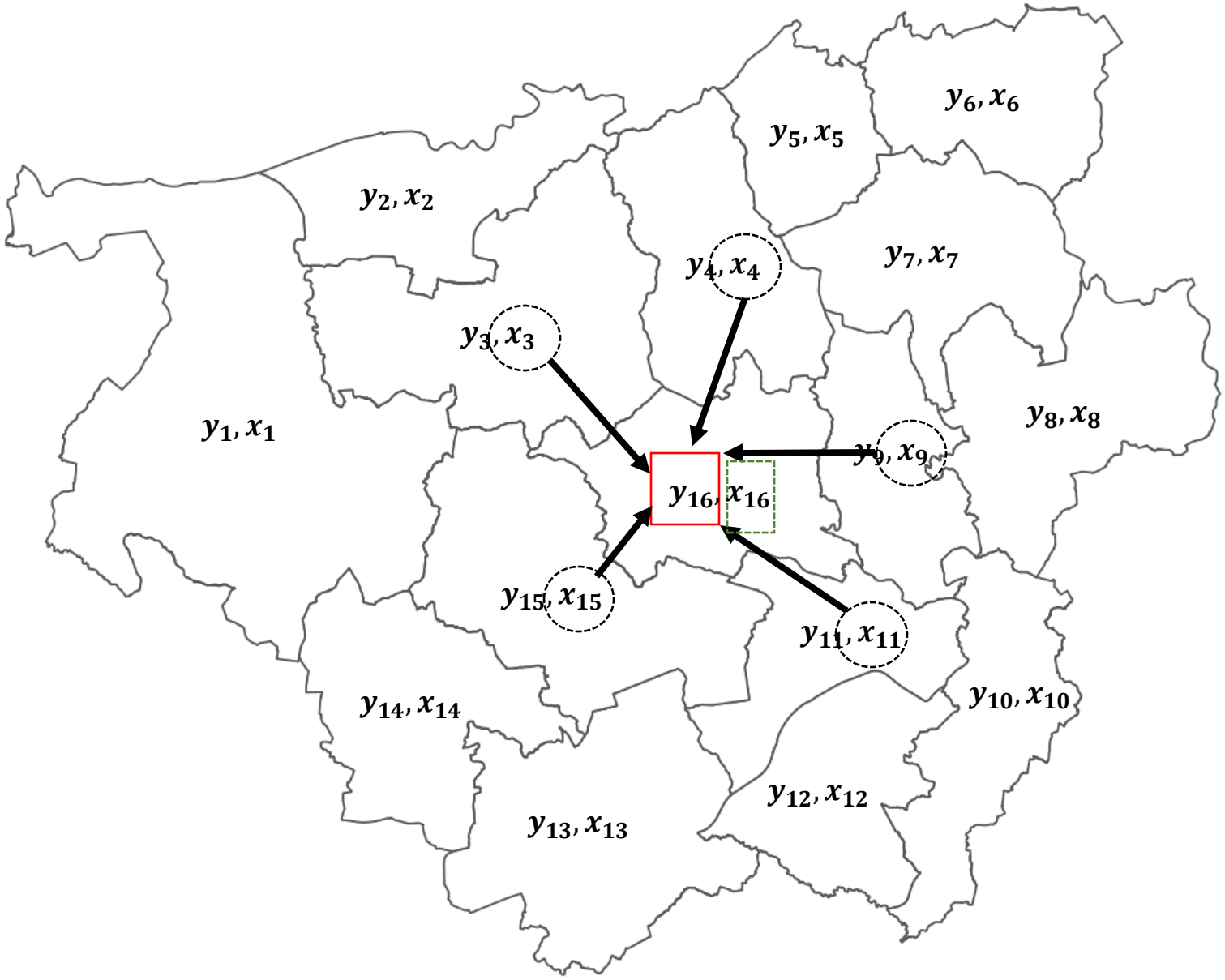
Notes:

The same concept applies to all other areas, as well. While the house price values in neighbouring areas from y_3, y_4 , y_9, y_{11} and y_{15} are impacting influence it on y_{16}.

We can say the same thing is happening for all other areas. For instances, let consider y_9. This is going to be influenced by other house price values from its direct neighbours (i.e., y_4, y_7 , y_8, y_{10} and y_{11}).

To reinforce, in this scenario, when the dependent variable's outcomes from neighbouring areas have an influence on the outcome of a location, we call this type of spatial dependence a **Spatial Lag on the dependent variable (aka Spatially lagged Y model)**

Scenario 2: Independent variable from the surrounding neighbourhoods have an influence on the values of the dependent variable



A hypothetical study area with 16 areas

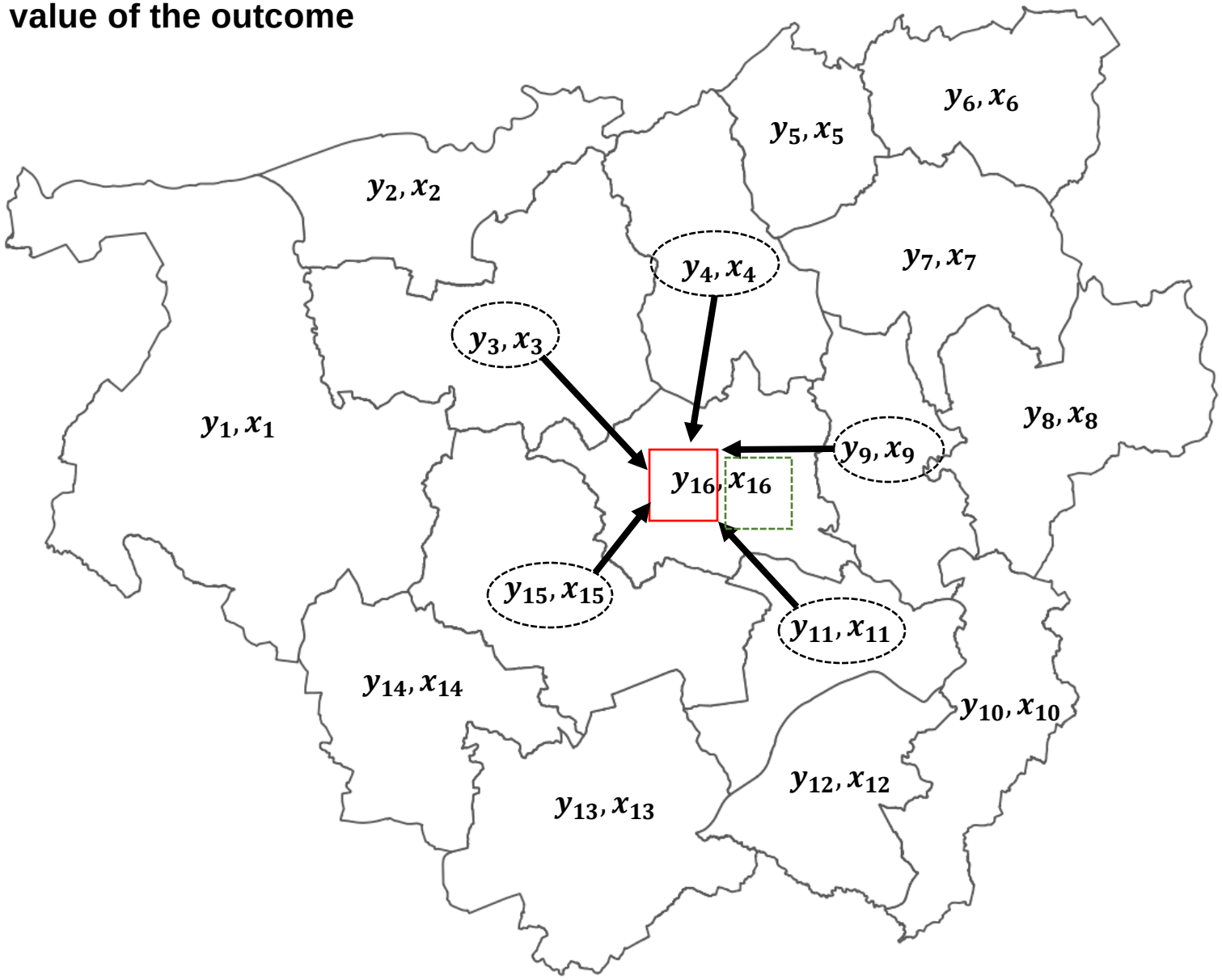
Notes:

Another type of spatial dependency in this neighbourhood structure can be seen when a **dependent variable (house price)** for some neighbourhood (16) (y_{16}) is influenced by the independent variable values (i.e., socioeconomic deprivation) from neighbouring areas (i.e., socioeconomic deprivation values x_3, x_4 , x_9, x_{11} and x_{15}) measured in neighbourhoods 3, 4, 9, 11, and 16.

In this scenario, when the independent variable from neighbours influences the outcome of a location, we call this type of spatial dependence a **Spatial Lag on the independent variable (aka Spatially lagged X model)**

Again, such spatial lag of this sorts are suggestive of spatial spill overs or diffusion.

Scenario 3: Both the values from the dependent and independent variable from the surrounding neighbourhoods have an influence on the value of the outcome



A hypothetical study area with 16 areas

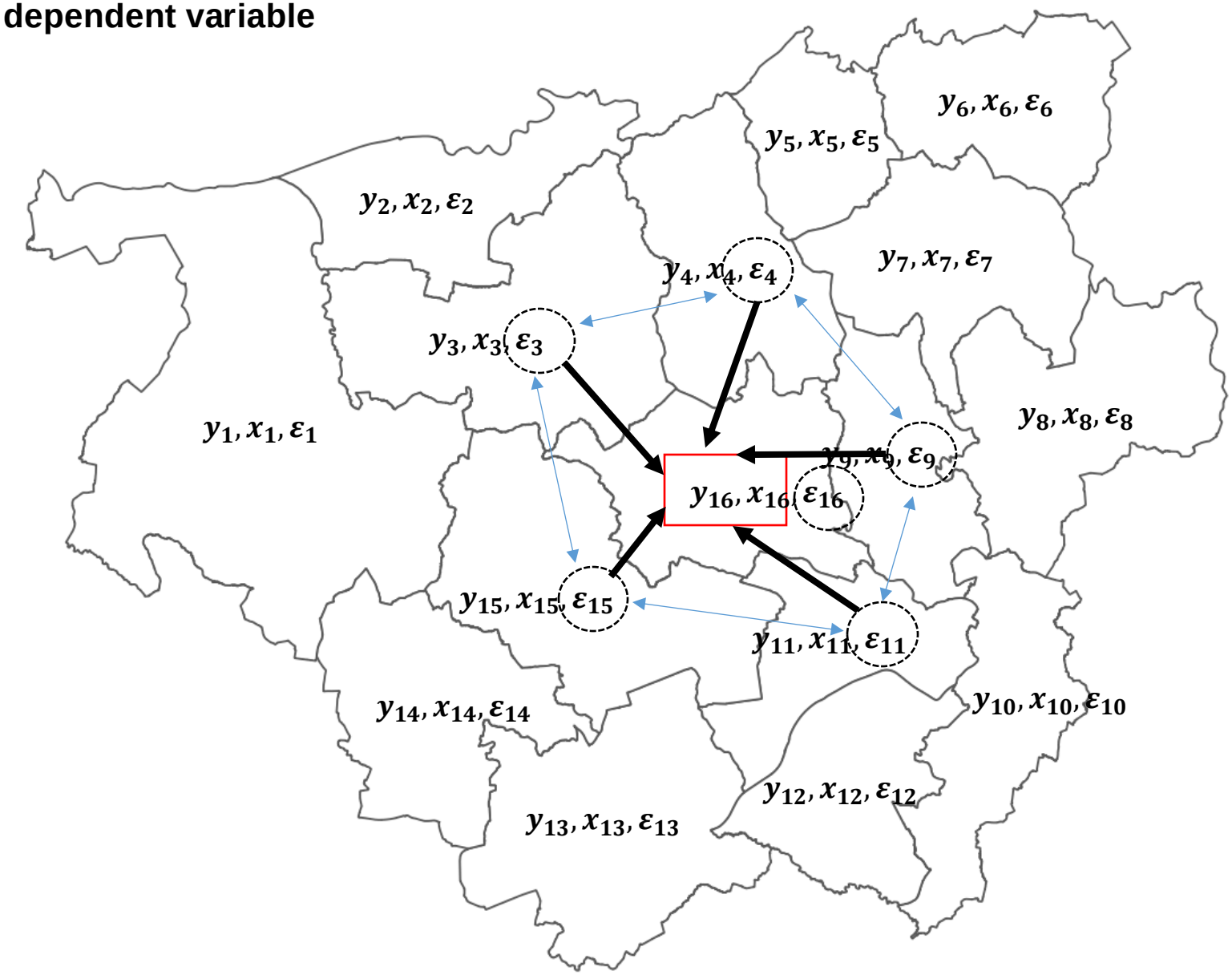
Notes:

The 3rd type of spatial dependency in this neighbourhood structure can be seen when a **dependent variable (house price)** for some neighbourhood (16) (y_{16}) is influenced by **BOTH** the values from the dependent and independent variable (i.e., house price and socioeconomic deprivation) from the surrounding neighbouring areas (i.e., socioeconomic deprivation values x_3, x_4 , x_9, x_{11} and x_{15} as well as house prices y_3, y_4 , y_9, y_{11} and y_{15}) measured in neighbourhoods 3, 4, 9, 11, and 16.

In this scenario, when both the dependent and independent variable from neighbours influences the outcome of a location, we call this type of spatial dependence a **Spatial Lag on the independent variable**

Again, such spatial lag of this sorts are suggestive of spatial spill overs or diffusion.

Scenario 4: The errors represent the unexplained or "independent variables" that are unaccounted across neighbourhoods impacting our dependent variable



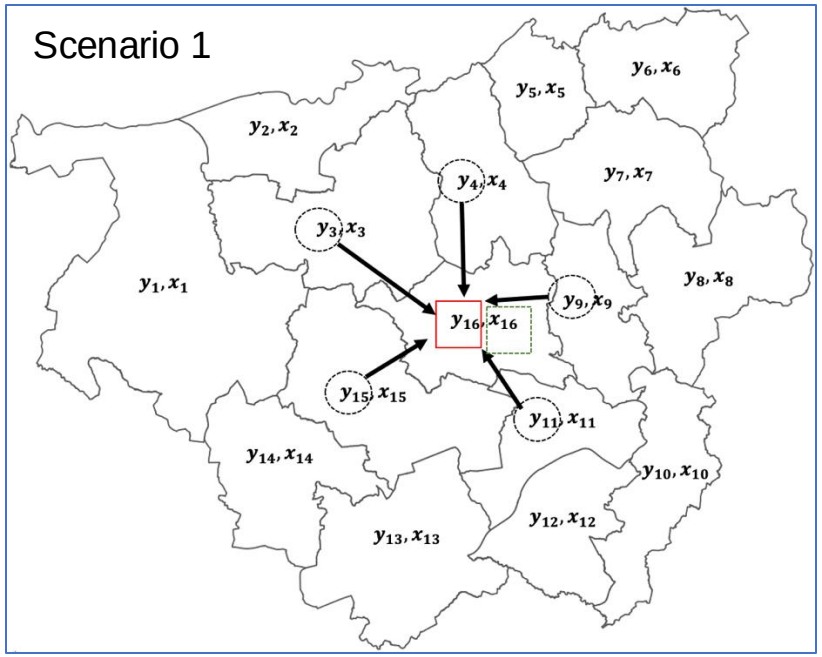
A hypothetical study area with 16 areas

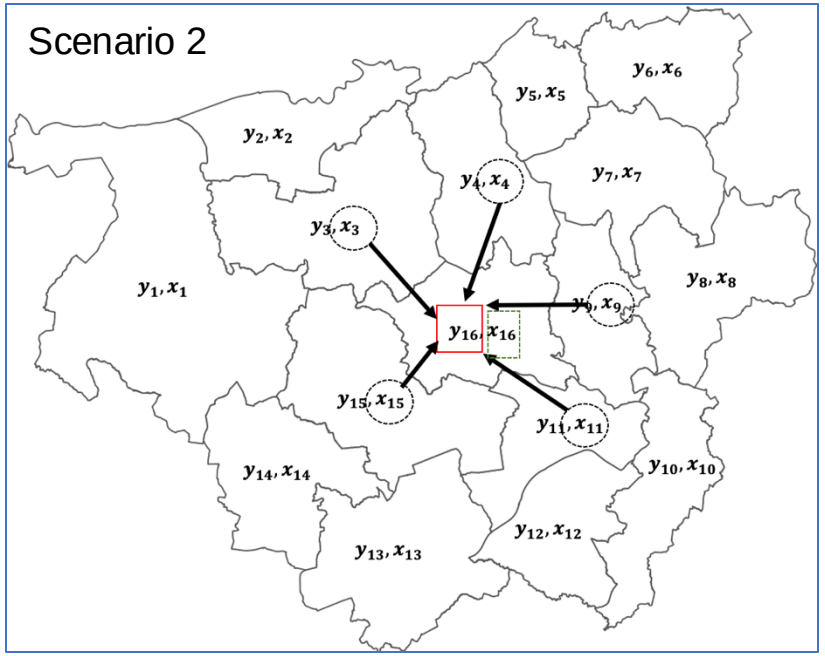
Notes:

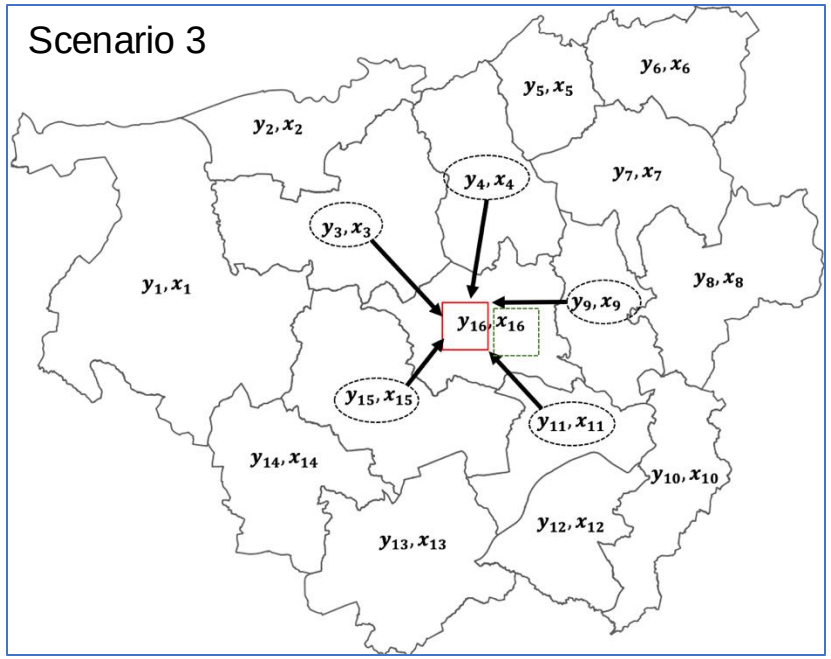
The last type of spatial dependency can occur when **the error** in the observed **dependent variable (house price)** from neighbouring areas where there's some independent variable which we have not considered which is somehow having a major influence on the measured outcome in neighbourhood (16) (y_{16}).

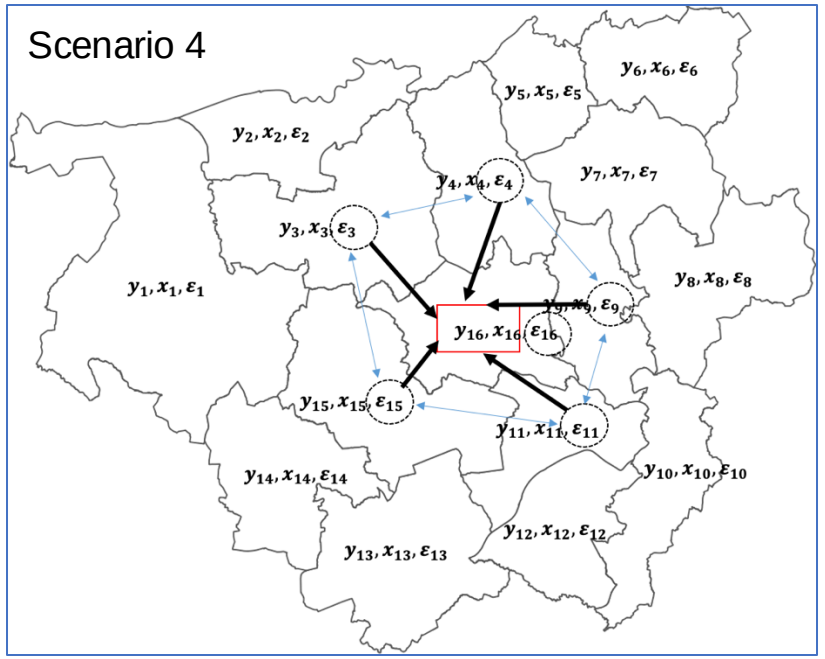
Errors, in turn, can also influence each other, and can be correlated.

In this scenario, this type of spatial dependence is called **Spatial Error**









Notes:

For these scenarios, where Y is the outcome, X are some independent variable(s) and E are residuals.

It possible to model the overall association, or relationships between X and Y, while considering:

1. Spatial configuration or layout of the for the study area.

2. The influences not only caused by the independent variables but also influences among the neighbouring outcome and error terms.

3. **<u>As well as quantitative the direct and indirectly relationship an independent variable will have on the outcome in a location, and in outcomes from neighbouring locations.</u>**

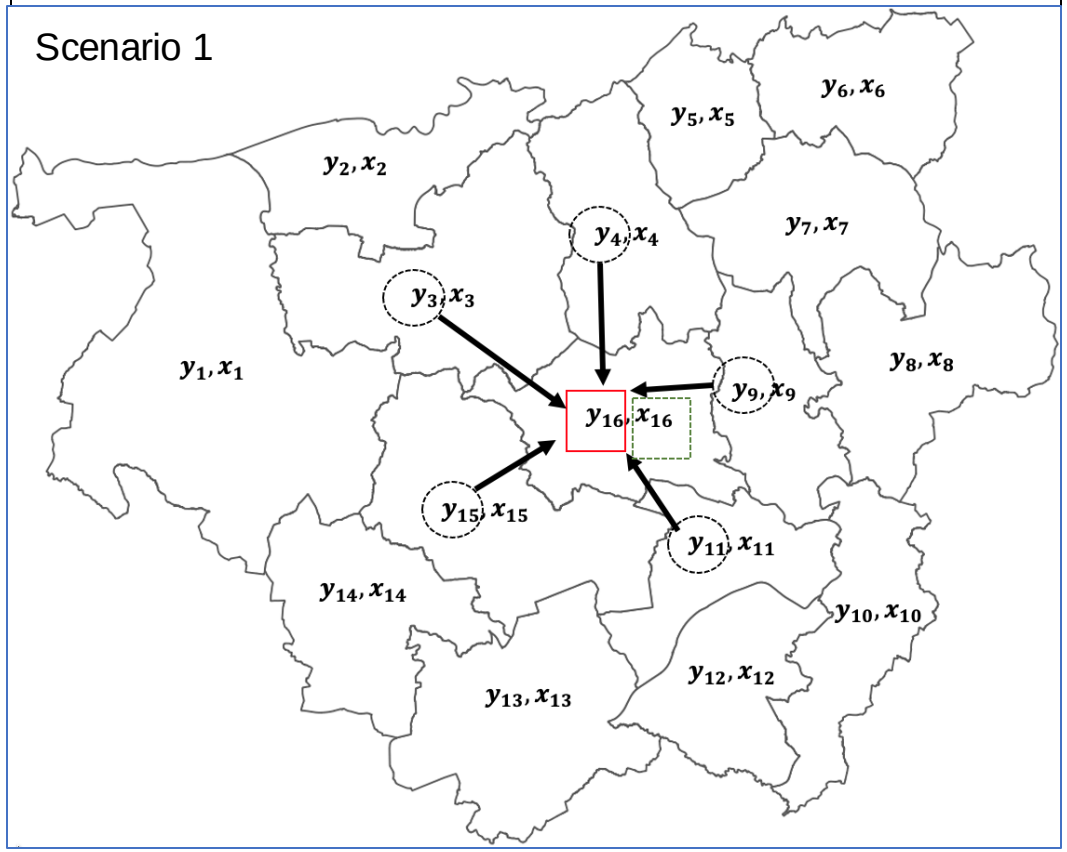
The linear regression can be extended into 4 different types of Spatial regression models which can feature a lag or error component.

Multivariable Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

1. Spatial Lag Model (lagged on the dependent variable)

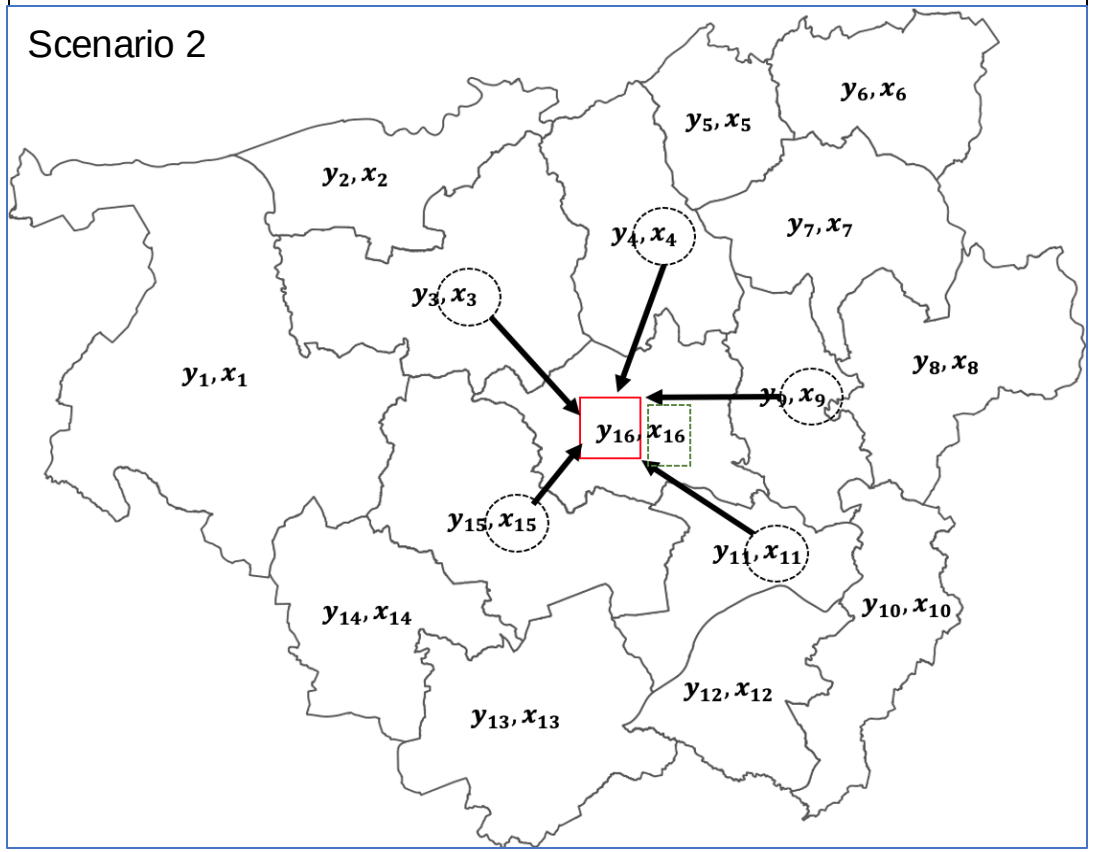
$$y = \rho WY + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$



- W is a spatial weights matrix (contiguity – based)
- Y in the righthand side of the equation represents the observed outcome from other neighbouring areas (i.e., spatially lagged on Y)
- ρ "Rho" is the degree for how each measure for the outcome of (**y)** is influenced by its direct neighbours (for **Y)**.

2. Spatial Lag Model (lagged on the independent variable)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \theta WX + \varepsilon$$



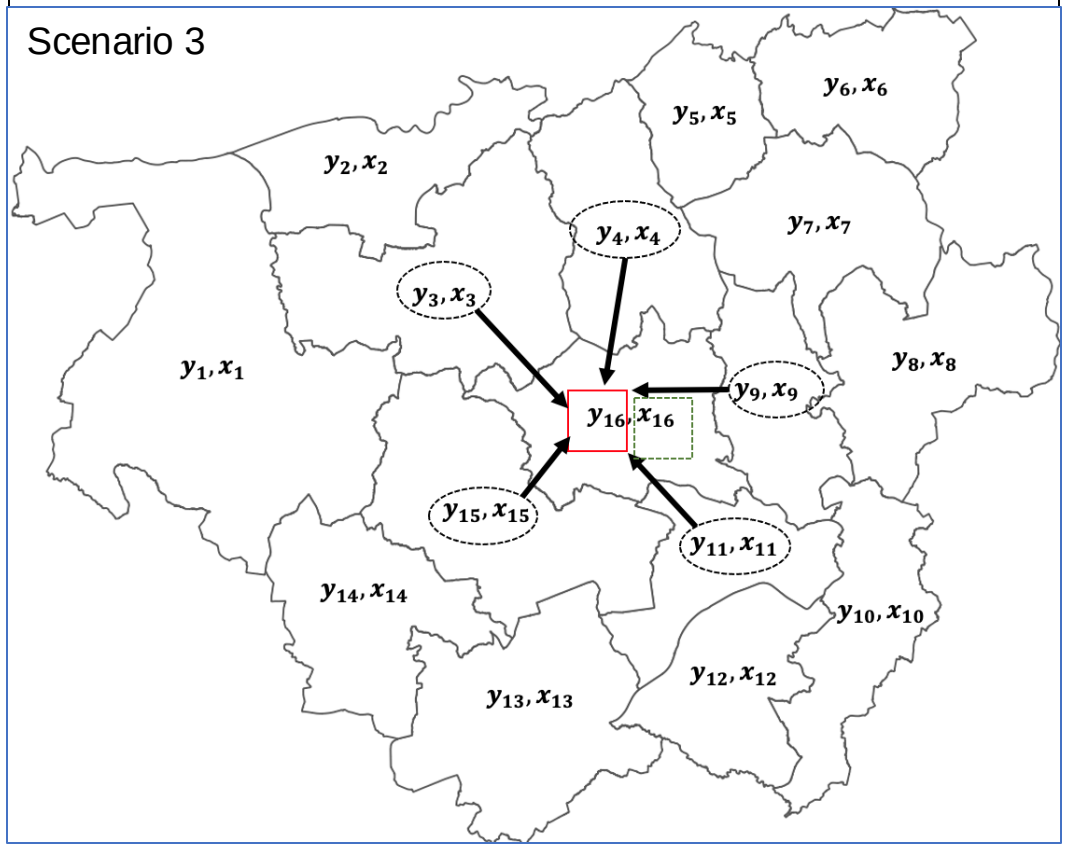
- W is a spatial weights matrix (contiguity – based)
- X in the righthand side of the equation represents the observed values of the independent variable from other neighbouring areas (spatially lagged on X)
- θ "theta" is the degree of how each measure of the outcome (**y**) is influenced by its direct neighbours (for X) measures.

Multivariable Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

3. Spatial Lag Model (lagged on both the dependent and independent variable)

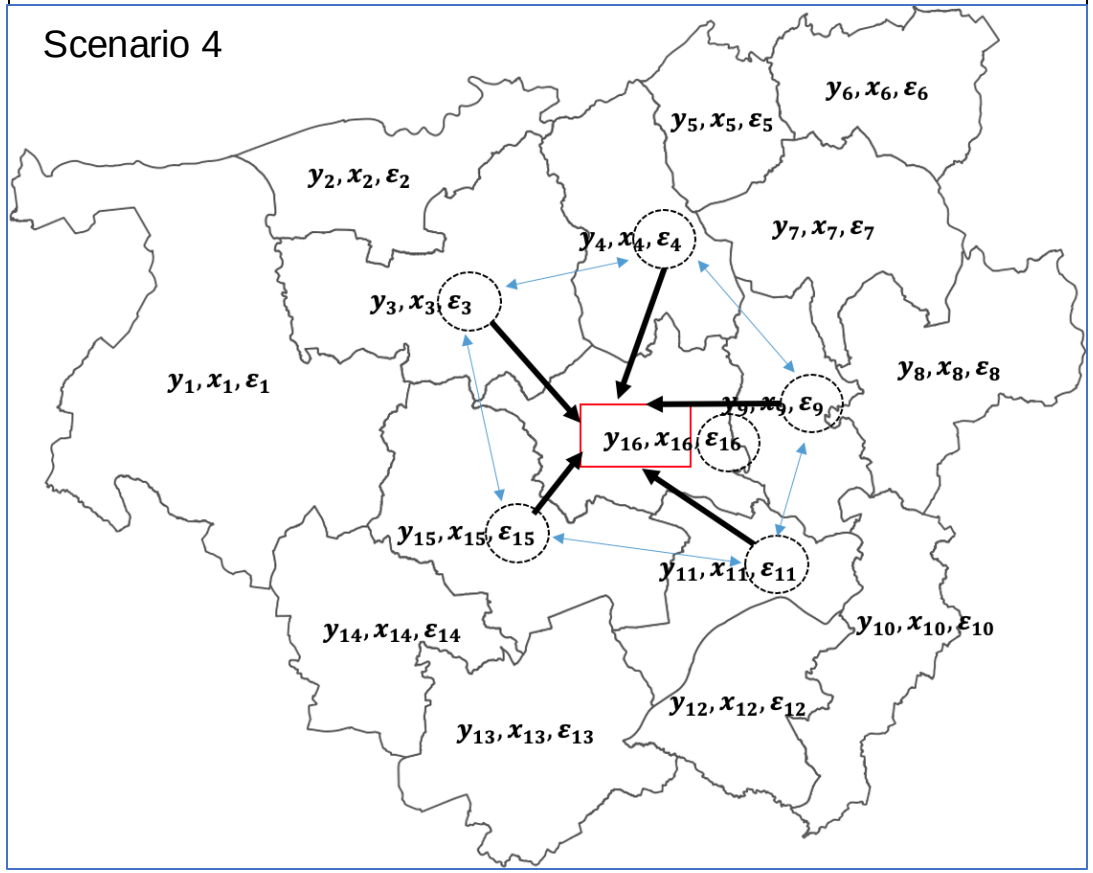
$$y = \rho WY + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \theta WX + \varepsilon$$



- Refer to slide number 28 to see the meanings of each parameters

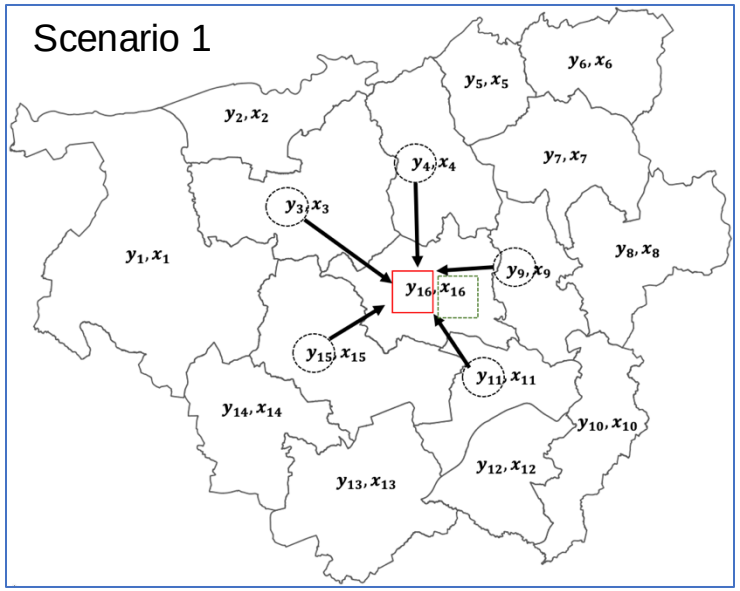
4. Spatial Error Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \lambda W\mathbf{u} + \varepsilon$$

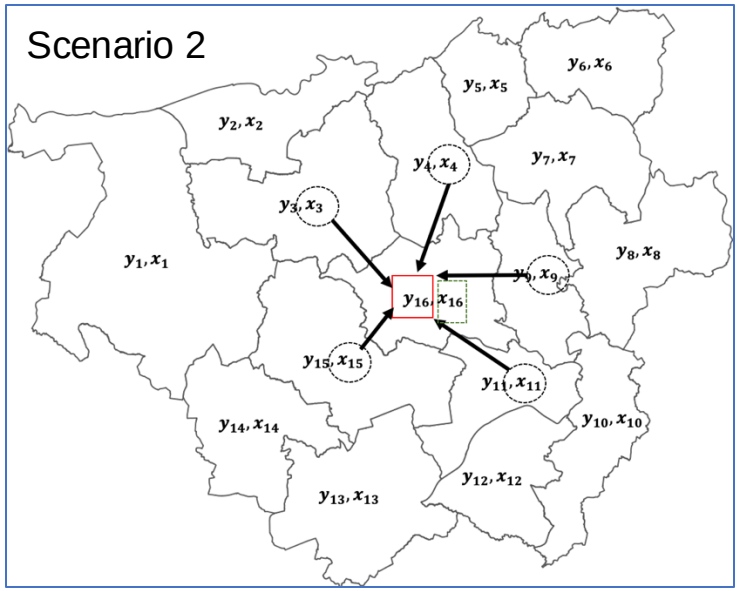


- u are the correlated spatial error terms
- λ "lambda" estimated coefficient for the product W and u.
- There is something spatial going on here that we have not accounted for!

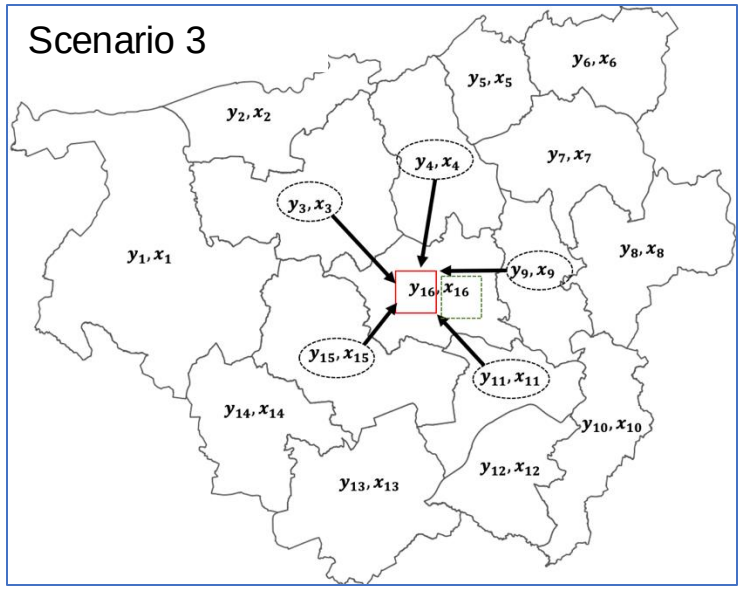
$$y = \rho WY + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$



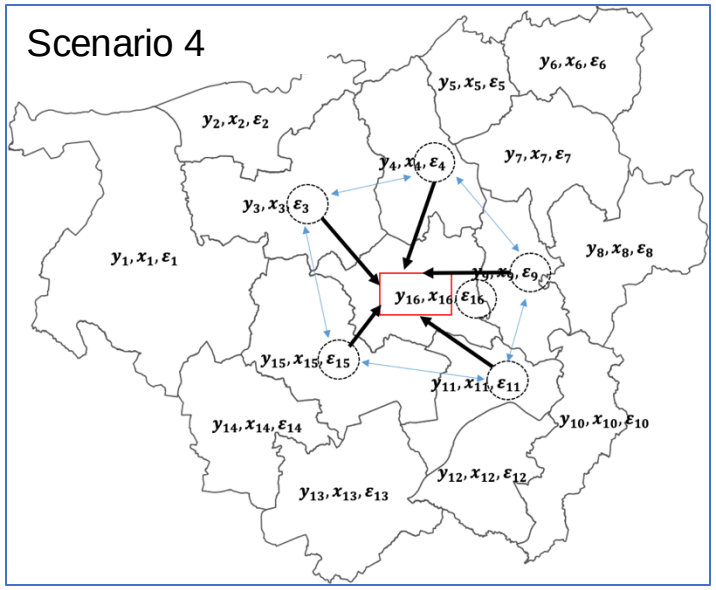
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \theta WX + \varepsilon$$



$$y = \rho WY + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \theta WX + \varepsilon$$



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \lambda W\mathbf{u} + \varepsilon$$



Notes:

For these scenarios, where Y is the outcome, X are some independent variable(s) and E are residuals.

It possible to model overall association or relationships between X and Y, while considering:

1. Spatial configuration or layout of the for the study area.

2. The influences not only caused by the independent variables but also influences among the neighbouring outcome and error terms.

The linear regression can be extended into 4 different types of Spatial regression models which can feature a lag or error component, or both.

Workflow for spatial regression modelling

Modelling process and model selection

When you want to conduct evidence-based analysis with spatial data – especially if the outcome is from a continuous distribution – you might want to follow these steps:

- **STEP 1: Carry some descriptive analysis to understand the underlying spatial distribution**
- **STEP 2: Perform a Linear regression in order to assess the residuals for the model output to determine whether not the assumptions of independence have been violated.**

If there's a violation:

- **STEP 3: Examine whether there is evidence of spatial dependence in the residuals using Moran's I test.**
- **If Moran's I test is not significant – STOP! NO EVIDENCE OF SPATIAL DEPENDENCE HENCE NO NEED OF SPATIAL APPROACH**

If Moran's I test is significant and positive – then map the residuals accordingly to examine the spatial patterning of the residual.

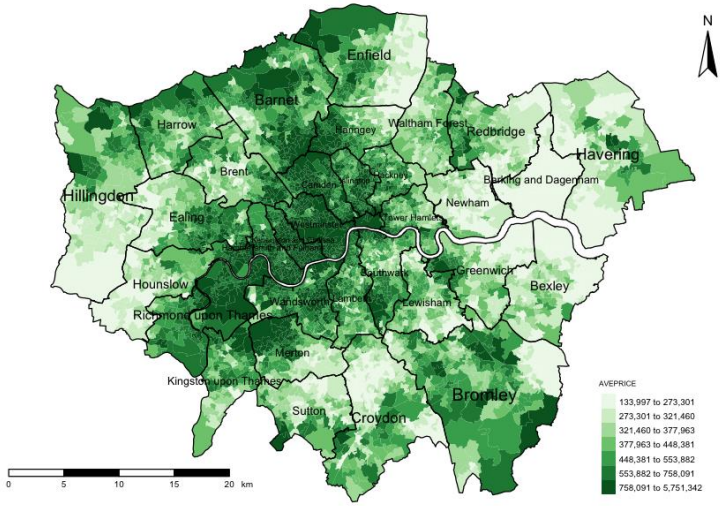
- **STEP 4: Pick one of the spatial regression model and run its regression. For Spatial Lag Y (or X) Model were interested interpreting the "rho" ("theta") estimates AND direct & indirect coefficients derived from the impacts() post estimation function.**
- **For the Spatial Error model, it's the "lambda" estimate. No need to use the impacts() function. We can get the coefficients directly from model.**
- **STEP 5: Compare the AIC to see which is a better model between the spatial regression(s) and linear regression too. The one with lowest is better.**
- **STEP 6: Perform the Moran's I test on the spatial model to see if they have adequately accounted for spatial autocorrelation. The model with a not significant p-value is the best! Select that model for making your causal or predictive inference**

If it's significant (i.e, Moran's I test) in step 6 – go for the one that has the lowest Moran's I statistic closest to zero.

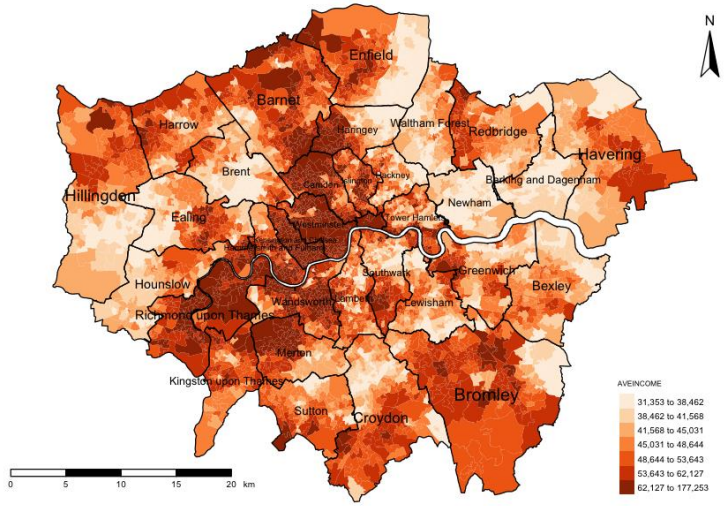
Example question: How does income, access to transportation and socioeconomic deprivation impact house price?

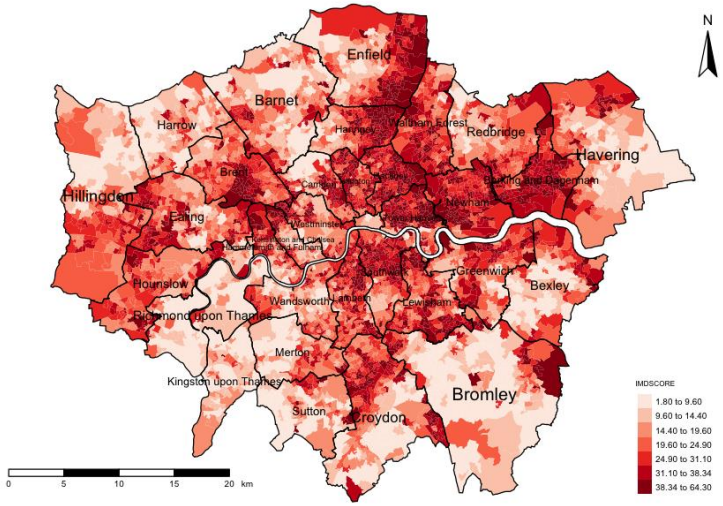
Step 1: Descriptive Analysis

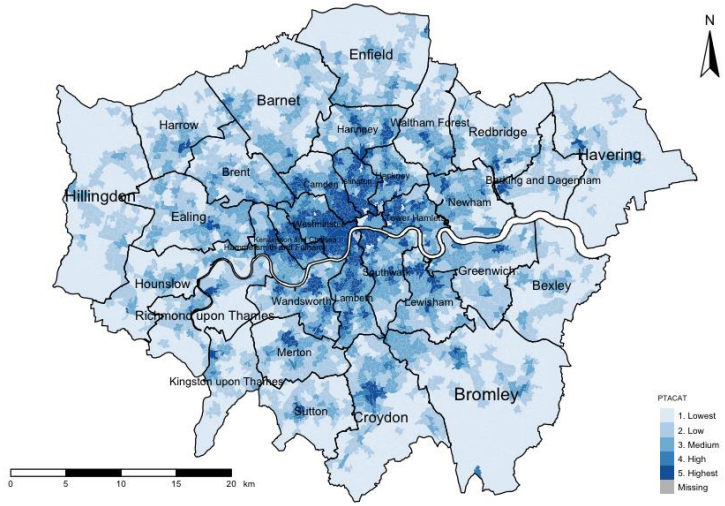
Average House Price (£) (Outcome)



Average Income (£) (Independent variable)







Socioeconomic deprivation score (Independent variable)

Public Transport Accessibility score (Independent variable)

Fitting a non-spatial model on spatial data for extracting the residuals

Step 2: Linear Regression Model

Variables	Coefficients	P-value
Intercept	-4.10	< 0.05
log(IMDSCORE)	0.136	<0.05
log(Income)	2.036	<0.05
log(PTA Index)	0.032	<0.05

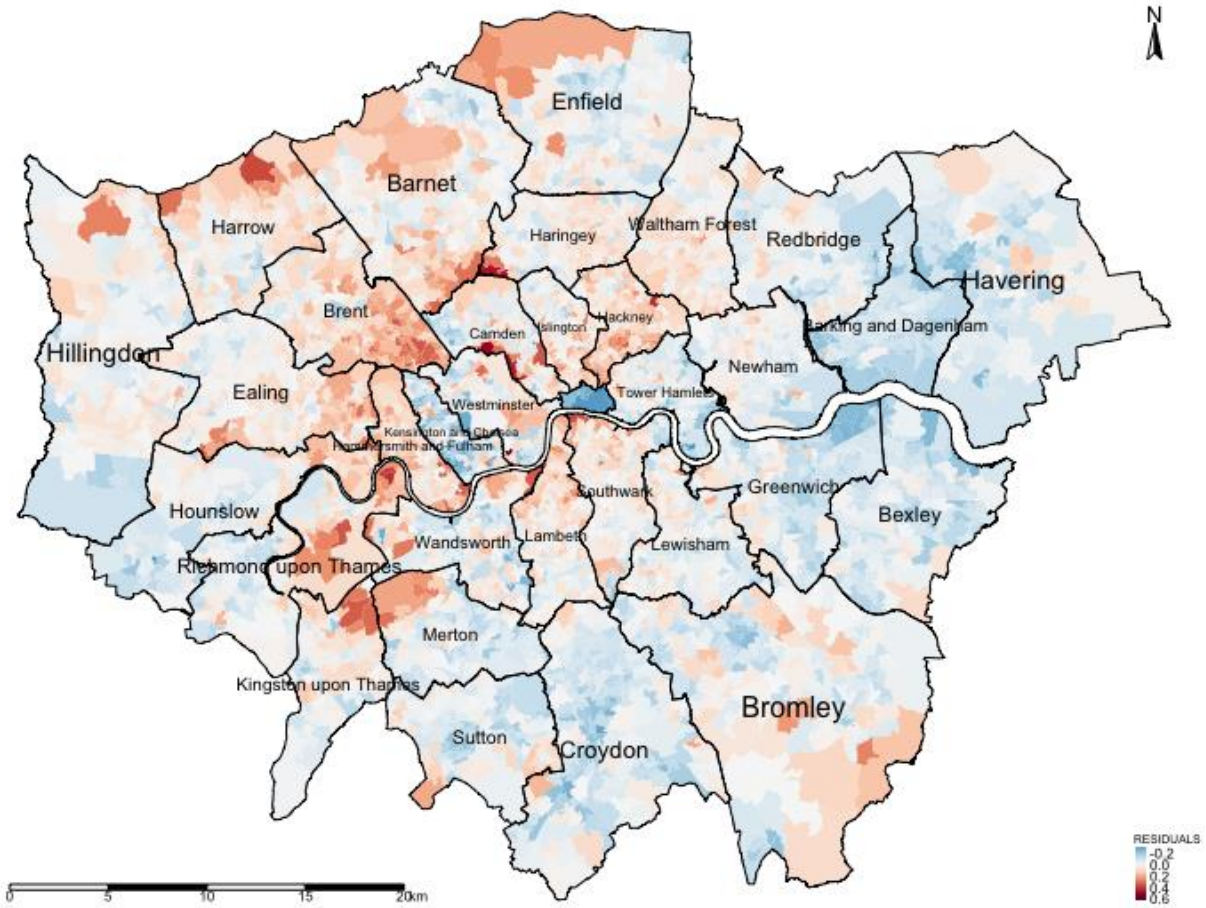
Explained variation in the model: 78.89%

Interpretation:

- On a log transformed scale, if the INCOME was to increase by 1.0%, we expect the house prices to increase by 2.04% and this average increase is statistically significant since the p-value = 0.00000000002 <0.05.
- On a log transformed scale, if the PTAINDEX was to increase by 1.0%, we expect the house prices to increase marginally by 0.03%. The marginally increase is statistically significant since the p-value = 0.000000000471 < 0.05.
- On a log transformed scale, if the IMDSCORE was to increase by 1.0%, we expect the house prices to increase marginally by 0.13% and this average increase is statistically significant since the p-value = 0.00000000002 <0.05.
- In terms of model performance: according to the Adjusted R-Squared value 0.7889 (78.89%)of the variation in the house prices across LSOAs were explained by the model after accounting for AVEINCOME, IMDSCORE, PTAINDEX. Since the Adjusted R-Squared more than 50.0% it is hence a very good model and significant (i.e., p-value = 0.0000000002 < 0.05).

Test for spatial dependency

Step 3: Plot the Residuals and Perform Moran's I Test



Interpretation of residuals for linear regression model

- Notice the spatial patterning and clusters of the LSOAs and the over-prediction (i.e., areas that have negative residuals, or blue tones) and under-prediction (i.e., areas that positive residuals, or red tones).
- This visual inspection of the residuals is telling you that spatial autocorrelation may be present here. This, however, would require a more formal test.

Moran's I = 0.475; p-value < 0.05

This is an indication that errors/residuals are related and not independent. Spatial regression for this output is needed.

Scenario 1: Spatial Lag Model (lagged on the dependent variable)

$$y = \rho WY + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Step 4: Run a spatial regression model (lagged Y)

- Estimate the ρ coefficient for WY
- Estimate the impacts i.e., the direct and indirect effects

ρ = 0.4522; p-value < 0.05

Interpretation:

The ρ statistic informs us of how the neighbouring LSOA house price values affect the house price an index location (y). The ρ value is a positive value of 0.4522 which means the house price from neighbouring LSOAs have a positive manner, and it is statistically significant (i.e., p-value < 0.05).

We can see the AIC for the lag model is lower than the original linear regression model (i.e., Lag: -9863.3 vs LM: -8510.8) therefore the lag model is okay.

```
Call:lagsarlm(formula = log10(AVEPRICE) ~ log10(IMDSCORE) + log10(AVEINCOME) +
    log10(PTAINDEX), data = spatialdatafile, listw = Residual_WeightMatrix)

Residuals:
       Min         1Q     Median         3Q        Max
-0.3608275 -0.0540603 -0.0039772  0.0518209  0.6492007

Type: lag
Coefficients: (asymptotic standard errors)
                   Estimate Std. Error  z value        Pr(>|z|)
(Intercept)      -2.6649683  0.0924917 -28.8130       < 2.2e-16
log10(IMDSCORE)   0.0435882  0.0068588   6.3551 0.0000000002083
log10(AVEINCOME)  1.2144821  0.0286833  42.3412       < 2.2e-16
log10(PTAINDEX)   0.0106795  0.0041275   2.5874        0.009671

Rho: 0.4522, LR test value: 1354.5, p-value: < 2.22e-16
Asymptotic standard error: 0.012282
    z-value: 36.819, p-value: < 2.22e-16
Wald statistic: 1355.6, p-value: < 2.22e-16

Log likelihood: 4937.637 for lag model
ML residual variance (sigma squared): 0.0077091, (sigma: 0.087801)
Number of observations: 4968
Number of parameters estimated: 6
AIC: -9863.3, (AIC for lm: -8510.8)
LM test for residual autocorrelation
test value: 443.54, p-value: < 2.22e-16
```

Step 5: Derive the direct and indirect coefficient values from the model

Interpretation of the relationship between our independent variables and outcome from spatial regression model

	Direct	Indirect	Total
log(IMD score)	0.045% (p < 0.05)	0.034 (p < 0.05)	0.079% (p < 0.05)
log(Income)	1.267% (p < 0.05)	0.949 (p < 0.05)	2.217% (p < 0.05)
log(PTA Index)	0.011% (p < 0.05)	0.008 (p < 0.05)	0.019% (p < 0.05)
	At specific location	**At neighbouring locations**	**Combined**

Interpretation:-

Direct effect for deprivation: This means that for every unit increase in deprivation (on log-scale) in a location (LSOA) leads to an average increase in the house prices by 0.045% in the **same location,** while accounting for spatial feedback (i.e., house price) from neighbouring locations (LSOAs) in London.

Indirect effect for deprivation: This means that for every unit increase in deprivation (on log-scale) in a location (LSOA) leads to average increase in the house prices by 0.034% **in neighbouring locations,** again, which accounting for its spillover effect through the spatial areas (LSOAs) in London.

Any questions?