# 空間モデリング

古谷知之

## 授業概要

- 線形回帰モデル（重回帰モデル）
- 同時自己回帰モデル
- 条件付き自己回帰モデル
- 誤差項の空間的自己回帰モデル
- 空間的自己相関モデル
- 空間ダービンモデル
- 地理的加重回帰モデル

# 統計モデルの種類

| | 主な推定方法 | データ分布 | 回帰係数 |
|---|---|---|---|
| 線形回帰モデル<br>（単回帰・重回帰など） | 最小二乗法 | 正規分布 | 一変数に一つ |
| 一般化線形モデル | 最尤推定法 | 正規分布以外の分布も可能 | 一変数に一つ |
| 一般化線形混合モデル | 最尤推定法 | 正規分布以外の分布も可能 | 変数の個体差に応じて推定可能 |
| 階層ベイズモデル | ベイズ推定 | 正規分布以外の分布も可能 | 変数の個体差に応じて推定可能 |

# 重回帰分析

- 従属変数$y$と$k$個の独立変数$x_1, x_2, \ldots, x_k$に対する標本数が$n$個の重回帰モデルは以下のように記述できる$(i = 1, \ldots, n)$。

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}
$$

# 重回帰分析

- 次のようなベクトルと行列を用いて、

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- 次式のように簡略化できる

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

# 重回帰分析で検討すべきこと

- 回帰係数の値：偏回帰係数
- 説明変数の重要性：標準化偏回帰係数
- 回帰係数の統計的有意性：t検定
- 回帰係数の信頼度：信頼区間
- 予測への適用可能性：重相関係数、決定係数（寄与率）
- 外れ値の検出：残差解析
- モデル全体の統計的有意性：F検定
- 従属変数間の相関：多重共線性
- 変数の選択：自由度修正済み決定係数

# 重回帰分析

- 誤差項$\boldsymbol{\varepsilon} = \boldsymbol{y} - X\boldsymbol{\beta}$の二乗和$Q$は、
$$Q = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T X^T X\boldsymbol{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$$
- 最小二乗法より、
$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\beta} = 0$$
- ここから以下の正規方程式を得る

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$$

- 両辺に左から$(X^T X)^{-1}$をかけると、回帰係数の推定量$\widehat{\boldsymbol{\beta}}$を得る

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T\boldsymbol{y}$$

# 線形回帰分析を行う上での仮定（前提）

- 線形回帰分析では、独立変数と従属変数がともに正規分布に従うことを前提としている

- 独立変数行列$X$が平均$\mu$、分散$\Sigma$の正規分布に従う$X \sim N(\mu, \Sigma)$とき、$X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\beta}\Sigma\boldsymbol{\beta}^T)$となる

- さらに誤差項$\boldsymbol{\varepsilon}$が平均$0$、分散$\sigma^2$の正規分布に従う$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$と仮定している。
- このことから従属変数$\boldsymbol{y}$は平均$X\boldsymbol{\beta}$、分散$\sigma^2 I$の正規分布に従う

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$$

# 重回帰モデルの統計量

- 回帰係数$\boldsymbol{\beta}$・誤差項$\boldsymbol{\varepsilon}$・従属変数$\boldsymbol{y}$の確率分布から、偏回帰係数$\widehat{\boldsymbol{\beta}}$・予測値$\widehat{\boldsymbol{y}}$・予測誤差$\boldsymbol{e}$の確率分布は以下のようになる

- 偏回帰係数

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

- 予測値

$$\widehat{\boldsymbol{y}} = X\widehat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \boldsymbol{y} = H\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 H)$$

$$H = X(X^T X)^{-1} X^T$$ $H$はハット行列

- 予測誤差

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (I - H)\boldsymbol{y} \sim N\left(0, \sigma^2 (I - H)\right)$$

# 重回帰モデルの統計量

- 偏回帰係数は正規分布に従う

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

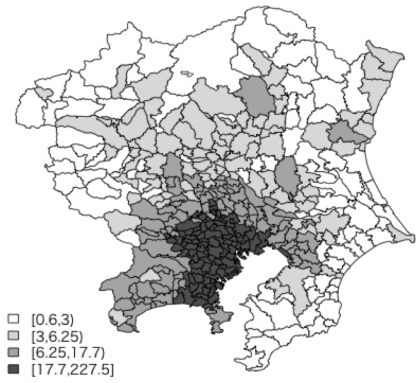
- この性質を標準化すると、以下のようになる

$$\frac{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\sigma^2 (X^T X)^{-1}}} \sim N(0, 1)$$

# 住宅地地価の重回帰モデル

- 関東圏の市区町村別住宅地地価（万円/㎡）は右図のように分布する
- 市区町村別住宅地地価を被説名変数、夜間人口密度（千人/㎡）と第三次産業従業人口密度（千人/㎡）を説明変数とする重回帰モデルを推定する

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

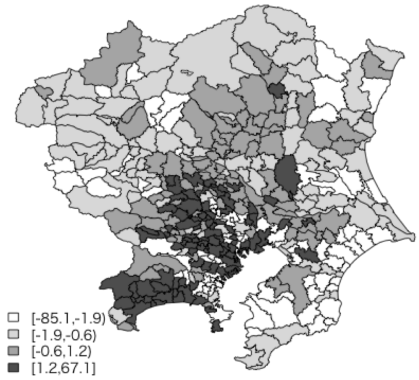



市区町村別住宅地地価の分布（万円/㎡）

# 住宅地地価の重回帰モデル

- 重回帰モデルの推定結果は右表のように得られた
- 説明変数の回帰係数は正でともに5%水準で統計的に有意であり、自由度修正済み$R^2$が0.83程度と高いことから、モデル推定結果は良好
- しかし誤差項は右図のように分布している

| 変数 | 回帰係数 | $t$ 値 |
|---|---|---|
| 定数項 | 2.55 | 4.50 |
| 夜間人口密度 | 1.68 | 15.85 |
| 第三次産業従業人口密度 | 2.25 | 28.65 |
| 自由度修正済み $R^2$ 値 | 0.828 | |




誤差項の分布

# 空間従属性

- 観測データや誤差項の空間的自己相関がある場合、用いるデータには空間的従属性がある可能性がある
- Moran's Iなどの指標を用いて空間的自己相関の有無を把握
- 被説明変数について空間従属性がある場合、隣接地域への空間的波及効果がある可能性がある
- 観測データ間で空間的自己相関がある場合、単純な対数尤度関数の和の最大化問題として、パラメータ推定を行うことは出来ないが、空間的自己相関を明示的に取り込むことで最尤推定法を適用できる
- 誤差項$\varepsilon$に空間的自己相関が存在すると考えられる場合、推定されたパラメータが統計的に有意であっても、それは見かけ上の相関にすぎず、特に通常最小二乗法を用いて推定されたパラメータは、一致性も不偏性ももたない。

# 空間的異質性(Spatial heterogeneity)

- 通常の線形回帰モデルでは、誤差項が独立に同一の正規分布に従うとしている
- つまり、誤差項の分散均一$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$を強く仮定している
- しかし、空間データはしばしば地域的・地理的な特徴を有するため、誤差項の分散が局地的に異なる可能性がある
- このような場合、誤差項の分散不均一（$\varepsilon_i \sim N\left(0, \sigma_i^2\right)$）を前提としたモデリングが必要となる
- 説明変数と被説明変数との関係が地域ごとに異なると考えられる場合、回帰係数$\boldsymbol{\beta}$の空間的異質性を考慮したモデル（地域ごとに回帰係数を推定する）の適用が望ましい

# 可変集計単位問題

- 規模の問題
  - 集計単位の規模を小さくすることにより、政策課題を空間的にきめ細かく検討できると期待できるかもしれない
  - しかしながら、局地的な空間的自己相関の発生や、小地域ではデータが観測されない、個人情報保護の観点からデータが公開されないといった、小地域統計独自の課題も有する

- ゾーニングの問題
  - 集計規模が同じであっても、地区の形状によって観測データの集計値が変化することがある
  - 行政境界を用いて分析を行う際には、規模の問題とゾーニングの問題が同時に発生するため、分析単位を適切に設定することが求められる

# 同時自己回帰モデル

- ある地域の誤差項が、自地域を含む他の地域の誤差項との相関関係を取り入れたモデル(simultaneous autoregressive model )
- 上述の地価モデルの場合、以下のように定式化できる

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$
$$\varepsilon_i = \sum_{i=1}^{n} b\varepsilon_i + e_i$$

- $\varepsilon_i$は空間的な系列相関を持つ誤差項、 $e_i$は空間的な系列相関を持たない誤差項

# 条件付き自己回帰モデル(CAR)

- 誤差項$\varepsilon_i$が地域$i$を除く誤差項に条件付けられることを明示したモデル(conditional autoregressive model)
- 地域$j$周辺の誤差項を選び$\varepsilon_{j\sim i}$とすると、その条件付き分布は次式のように表すことができる

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$
$$\varepsilon_i | \varepsilon_{j\sim i} = N\left(\sum_{j\sim i} \frac{c_{ij}\varepsilon_i}{\sum_{j\sim i} c_{ij}}, \frac{\sigma^2_{\varepsilon_i}}{\sum_{j\sim i} c_{ij}}\right)$$

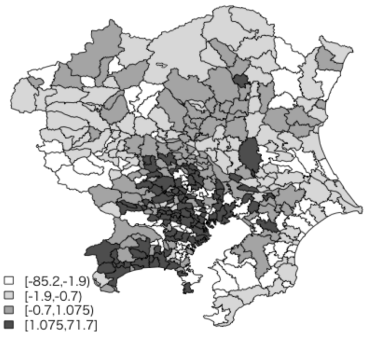
- 共分散行列$V$は次式のようになる

$$V = (I - \rho W)^{-1}\Sigma$$

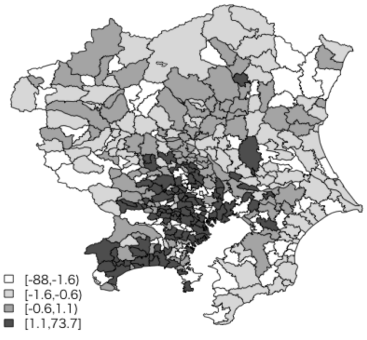
# SARとCARの推定結果の比較

| | SAR | | CAR | |
|---|---|---|---|---|
| 変数 | 回帰係数 | Z値 | 回帰係数 | Z値 |
| 定数項 | 3.00 | 4.26 | 3.23 | 4.45 |
| 夜間人口密度(POPD) | 1.58 | 12.73 | 1.54 | 12.19 |
| 第三次産業従業人口密度(EMP3D) | 2.18 | 25.31 | 2.15 | 24.82 |
| λ | 0.22 | 4.54[1)] | 0.44 | 4.53[1)] |
| AIC | 2552.1 | | 2552.1 | |

1) Z値ではなくLR検定統計量




SARの誤差項分布




CARの誤差項分布

# 空間的自己回帰(SAR)モデル

- 被説明変数に空間的従属性を表現したモデル(spatial auto-regression model)
- 空間的波及効果を定式化したモデル

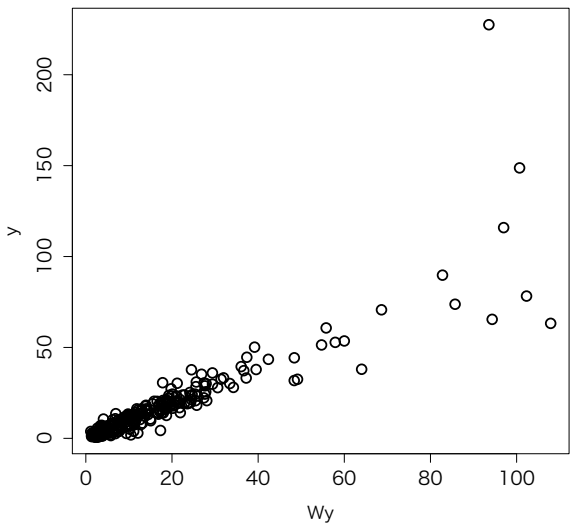
$$\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

- 共分散行列$V$は次式のようになる

$$V = (I - \rho W)^{-1}\Sigma(I - \rho W')^{-1}$$

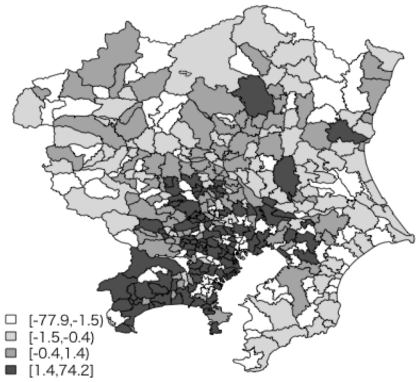
# $\boldsymbol{y}$と$W\boldsymbol{y}$との相関（地価モデルの場合）

- 地価モデルについて$\boldsymbol{y}$と$W\boldsymbol{y}$との相関を取ると、右図のようになる
- 両者の間には相関があるように見える



# 空間的自己回帰(SAR)モデルの推定結果

| | 最尤推定法 | | 二段階最小二乗法 | |
|---|---|---|---|---|
| 変数 | 回帰係数 | $Z$ 値 | 回帰係数 | $t$ 値 |
| 定数項 | 1.71 | 4.50 | 1.27 | 2.36 |
| 夜間人口密度(POPD) | 0.71 | 5.06 | 0.20 | 1.01 |
| 第三次産業従業人口密度(EMP3D) | 1.60 | 15.49 | 1.26 | 9.18 |
| $\rho$ | 0.44 | 8.41 | 0.67 | 8.37 |
| | AIC=2504 | | 残差分散=55.47 | |




誤差項分布

# 誤差項の空間的自己回帰モデル(SEM)

- 誤差項の空間的自己回帰モデル(spatial error model)は誤差項に空間的自己相関を明示したモデル

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$$
$$\boldsymbol{u} = \lambda W u + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

- 共分散行列$V$は次式のようになる

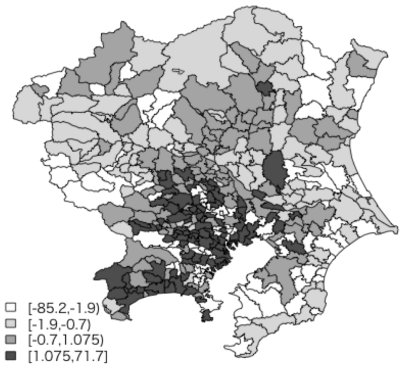
$$V = (I - \lambda W)^{-1}\Sigma(I - \lambda W')^{-1}$$

- 最尤法の他に、一般化モーメント法（$\lambda$と$\sigma^2$を同時に最適化する方法）、ベイズ法などを用いて推定される

# SEMの推定結果

| | 最尤推定法 | | 一般化モーメント法 | |
|---|---|---|---|---|
| 変数 | 回帰係数 | Z値 | 回帰係数 | Z値 |
| 定数項 | 3.00 | 4.26 | 2.81 | 4.34 |
| 夜間人口密度(POPD) | 1.58 | 12.73 | 1.63 | 13.90 |
| 第三次産業従業人口密度(EMP3D) | 2.18 | 25.31 | 2.21 | 26.56 |
| λ | 0.22 | 2.73[1)] | 0.14 | 3.99[1)] |
| | AIC=2552.1 | | AIC=2552.6 | |

1)Z値ではなくLR検定統計量を示している




誤差項分布

# 空間ダービンモデル

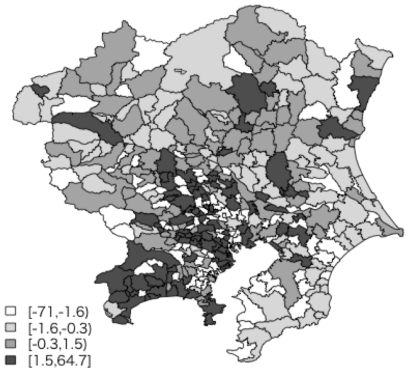
- 空間ダービンモデルは説明変数と被説明変数の両方に空間的従属性を取り入れたモデル

$$\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \lambda WX\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

# 空間ダービンモデルの推定結果

| 変数 | 回帰係数 | Z値 |
|---|---|---|
| 定数項 | 1.35 | 2.43 |
| 夜間人口密度(POPD) | -0.49 | -1.83 |
| 第三次産業従業人口密度(EMP3D) | 1.44 | 11.36 |
| $\rho$（$Wy$） | 0.21 | 2.57 |
| $\rho$（夜間人口密度） | 1.88 | 5.75 |
| $\rho$（第三次産業従業人口密度） | 0.63 | 2.42 |

AIC=2475.4（線形回帰モデルの AIC=2479.2）




誤差項分布

# 空間従属性のラグランジュ乗数検定

- 観測データや誤差項に空間的従属性を取り入れるかどうかについては、空間的従属性を取り入れなかった場合、すなわち通常の線形回帰モデルと比較して、空間的従属性を考慮することの統計的有意性を判断する方法が提案されている
- 空間従属性に関する回帰係数$[\rho, \lambda] = \theta$について、帰無仮説と対立仮説を次のようにおく

$$帰無仮説 H_0: \theta = [\rho, \lambda] = 0$$

$$対立仮説 H_1: \theta = [\rho, \lambda] \neq 0$$

- ラグランジュ乗数$LM$について$\chi^2$検定を行う（$L$は対数尤度）

$$LM = (\partial L / \partial \theta)$$

# 空間従属性のラグランジュ乗数検定

- 地価モデルのSARモデルとSEMモデルについてラグランジュ乗数検定をした結果は以下の表の通り

| モデル | ラグランジュ乗数（$LM$） | $p$ 値 |
|---|---|---|
| SAR モデル（LMlag） | 46.52 | $9.05\times10^{-12}$ |
| SEM モデル（LMerr） | 3.43 | 0.064 |

# マルチレベルモデル

- グループ$j$に属する地域$i$について、以下のような線形回帰モデルを考える

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim N\left(0, \sigma_y^2\right)$$

- ここで、誤差項は未知である
- 定数項と回帰係数について、グループ毎の違いを認めるかどうかによって、柔軟なモデル推定ができる

# マルチレベルモデル

① 定数項と回帰係数が変動しないモデル

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} + \varepsilon_{ij}$$

② 定数項のみが変動するモデル

$$y_{ij} = \beta_{j0} + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} + \varepsilon_{ij}$$

③ 回帰係数のみが変動するモデル

$$y_{ij} = \beta_0 + \beta_{j1} x_{ij1} + \cdots + \beta_{jk} x_{ijk} + \varepsilon_{ij}$$

④ 定数項と回帰係数の両方が変動するモデル

$$y_{ij} = \beta_{j0} + \beta_{j1} x_{ij1} + \cdots + \beta_{jk} x_{ijk} + \varepsilon_{ij}$$

# ランダム効果と固定効果

- ランダム効果
  - 回帰係数や定数項について、個人や地域・グループでの変動を認めるモデル
- 固定効果
  - 上記のような変動を認めないモデル
- 混合効果
  - ランダム効果と固定効果が混在したモデル

# マルチレベルモデル

- レベル 1 =lower level

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim N\left(0, \sigma_y^2\right)$$
$$y_{ij} \sim N\left(\beta_0 + \beta_1 x_{ij1}, \sigma_y^2\right)$$

- レベル 2 =upper level

$$\beta_{j0} = \gamma_{00} + \gamma_{10} u_{j0} + \eta_{j0}$$
$$\beta_{j1} = \gamma_{01} + \gamma_{11} u_{j0} + \eta_{j1}$$
$$\eta_{jk} \sim N\left(0, \sigma_\eta^2\right)$$
$$\beta_{jk} \sim N\left(0, \sigma_{\beta_k}^2\right)$$
$$k = \{0, 1\}$$

# マルチレベルモデル

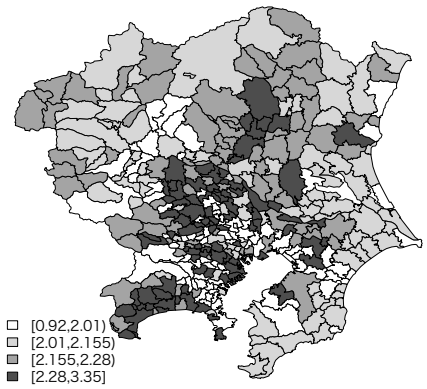
- レベル 1 とレベル 2 を合わせると、以下のように式変形される

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \varepsilon_{ij}$$
$$= \left(\gamma_{00} + \gamma_{10} u_{j0}\right) + \left(\gamma_{01} + \gamma_{11} u_{j0}\right) x_{ij1} + \eta_j + \varepsilon_{ij}$$
$$= \underbrace{\left(\gamma_{00} + \gamma_{01} x_{ij1}\right)}_{\text{固定効果}} + \underbrace{\left(\gamma_{10} u_{j0} + \gamma_{11} u_{j0} x_{ij1}\right)}_{\text{ランダム効果}} + \eta_j + \varepsilon_{ij}$$

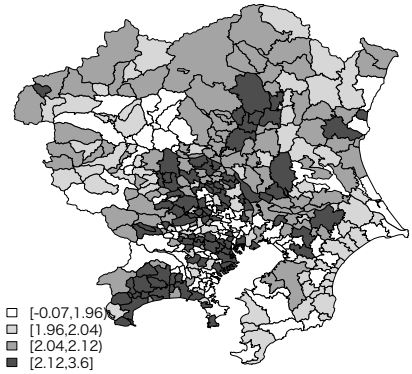
- この式は、次式のように整理できる

$$\boldsymbol{y_i} = \boldsymbol{X_i B} + \boldsymbol{Z_i b_i}$$
$$\boldsymbol{b_i} \sim N\left(0, \sigma_{b_i}^2\right)$$

# マルチレベルモデルの推定結果の例




夜間人口密度




第三次産業従業人口密度

# 地理的加重回帰(GWR)モデル

- 地理的加重回帰モデル（geographically weighted regression model: GWR）は空間的異質性と空間的従属性の両方を考慮した空間モデル
- 地域$i$ごとに異なる回帰係数$\beta_i$と空間的従属性を示す空間重み付け行列$W_i$とで構成される

$$W_i y = W_i X \beta_i + \varepsilon_i$$
$$\varepsilon_i = N(0, \sigma^2 V_i)$$

- ここで、 $V_i$は対角要素$\{v_i, \ldots, v_N\}$から構成される対角行列

# GWRの空間重み付け行列

- 空間重み付け行列$W_i$には、地域$i$と他地域との距離$d_i$及びバンド幅$\theta$を用いた、様々な関数が提案されている

- 指数関数

$$W_i = exp(-d_i/\theta)$$

- bri-cube関数

$$W_i = (1-(d_i/q_i)^3)^3$$

- ガウス関数

$$W_i = exp\left(-\frac{(d_i/\theta)^2}{2}\right)$$

- bi-square関数

$$W_i = exp(-(d_i/\theta)^2)$$

# GWRの空間重み付け行列

- 空間重み付け行列$W_i$は以下の対角行列で構成される

$$W_i = \begin{pmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & w_{iN} \end{pmatrix}$$

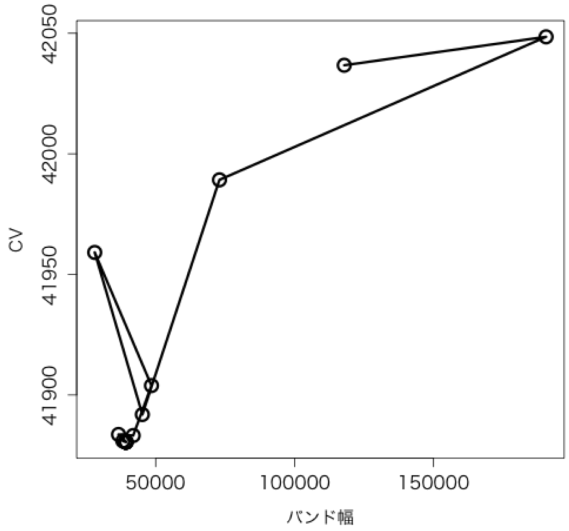
- 例えば指数関数を用いた場合、その要素$w_{ij}$は次式のような基準化された値を与える

$$w_{ij} = \frac{exp\left(-d_{ij}/\theta\right)}{\sum_{j=1}^{N} exp\left(-d_{ij}/\theta\right)}$$

# バンド幅の求め方

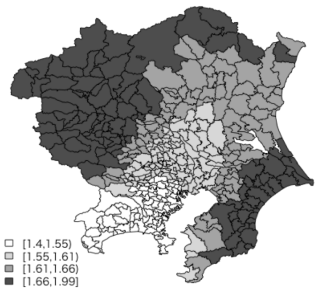
- バンド幅の計算には、交差検証（クロスバリデーション）法を用いる
- 次式で与えられるクロスバリデーションスコア$CV$を最小化する

$$CV = \sum_{i=1}^{N} [y_i - \hat{y}_{\neq i}(\theta)]$$

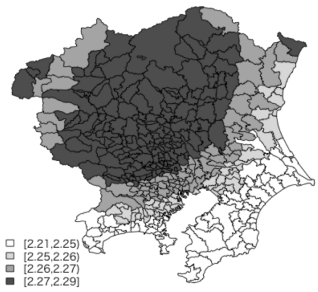



# GWRの推定結果の例

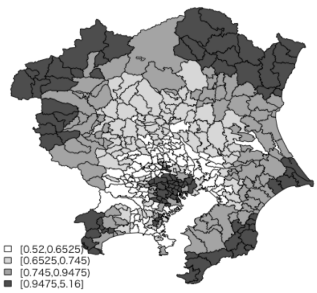
夜間人口密度




第三次産業
従業人口密度




標準誤差




ローカルな$R^2$値