# 統計解析

古谷知之

## 授業概要

* 履修者の状況に応じて変更される場合がありますが、全体としては以下のような授業構成となります。
* 講義の中でR演習を行うこともあります。

第1回　ガイダンス・単回帰分析
第2回　重回帰分析(1)
第3回　重回帰分析(2)
第4回　一般化線形回帰モデル(1)
第5回　一般化線形回帰モデル(2)
第6回　一般化線形回帰モデル(3)
第7回　一般化線形回帰モデル(4)
第8回　一般化線形回帰モデル(5)
第9回　一般化線形回帰モデル(6)
第10回　一般化線形混合モデル
第11回　状態空間モデル
第12回　R演習(1)
第13回　R演習(2)
第14回　R演習(3)

# 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 |

# 本授業で扱う統計モデル

- 線形回帰モデル
  - 単回帰モデル、重回帰モデル
- 一般化線形回帰モデル
  - 離散：ポアソン回帰モデル、二項反応モデル（ロジスティック回帰モデル、プロビット回帰モデル、補対数対数モデル）、負の二項分布モデル、ゼロ過剰ポアソン回帰モデル、ゼロ過剰負の二項分布モデル
  - 連続：ガンマ回帰モデル、ベータ回帰モデル、指数-ガウス回帰モデル
  - スパース：Lasso回帰モデル、Ridge回帰モデル
- 一般化線形混合モデル
  - マルチレベルモデル
- 状態空間モデル

# 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

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

- 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（混合効果モデル）

- 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

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

# マルチレベルモデル（混合効果モデル）

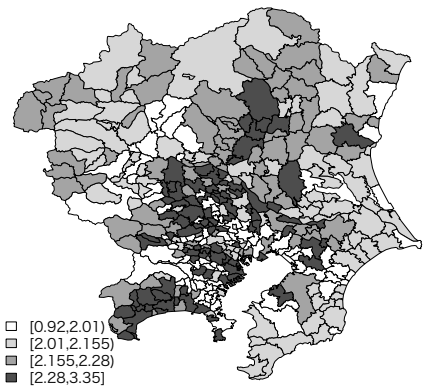
- レベル 1 とレベル 2 を合わせると、以下のように式変形される

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

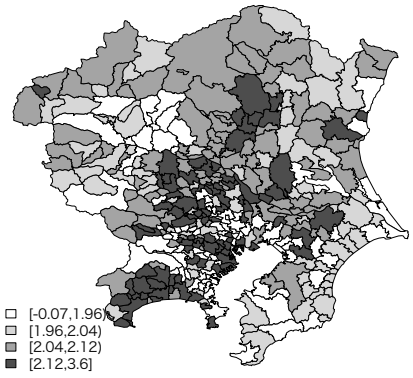
- この式は、次式のように整理できる

$$
\begin{aligned}
\boldsymbol{y_i} &= \boldsymbol{X_i B} + \boldsymbol{Z_i b_i} + \varepsilon_{ij} \\
\boldsymbol{b_i} &\sim N\left(0, \sigma_{b_i}^2\right) \\
\varepsilon_{ij} &\sim N\left(0, \sigma_y^2\right)
\end{aligned}
$$

# 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

# マルチレベルモデルの最尤推定

- 混合効果モデル

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

- $\boldsymbol{y_i}$の共分散行列$\boldsymbol{V_{y_i}}$は以下のようになる

$$\boldsymbol{V_{y_i}} = \boldsymbol{Z_i}\left(\sigma_{b_i}^2 I\right)\boldsymbol{Z_i}' + \sigma_y^2 I$$

- $\boldsymbol{y_i} \sim N\left(\boldsymbol{X_i B} + \boldsymbol{Z_i b_i}, \boldsymbol{V_{y_i}}\right)$および$\boldsymbol{b_i} \sim N\left(0, \sigma_{b_i}^2\right)$であることから、ランダム効果$\boldsymbol{Z_i b_i}$に対する期待値$E(\boldsymbol{Z_i b_i}) = 0$である
- したがって、 $\boldsymbol{y_i}$の周辺分布は$\boldsymbol{y_i} \sim N\left(\boldsymbol{X_i B}, \boldsymbol{V_{y_i}}\right)$となる

# マルチレベルモデルの最尤推定

- このとき、混合効果モデルの尤度関数$\ell$は次式で表せる

$$\ell = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{XB})'\boldsymbol{V_y}^{-1}(\boldsymbol{y} - \boldsymbol{XB})\right]}{(2\pi)^{(n/2)}\sqrt{\boldsymbol{V_y}}}$$

- 対数尤度関数$\log(\ell)$は次式のようになる

$$\log(\ell) = -\frac{\log(2\pi)}{2} - \frac{1}{2}\left[\log\left|\boldsymbol{V_y}\right| + (\boldsymbol{y} - \boldsymbol{XB})'\boldsymbol{V_y}^{-1}(\boldsymbol{y} - \boldsymbol{XB})\right]$$

# マルチレベルモデルの最尤推定

- 対数尤度関数$\log(\ell)$が最大となるとき、

$$\boldsymbol{X'V_y^{-1}(y - XB)} = 0$$

- 固定効果$\boldsymbol{B}$の最尤推定量$\widehat{\boldsymbol{B}}$

$$\widehat{\boldsymbol{B}} = \left(\boldsymbol{X'V_y^{-1}X}\right)^{-1}\boldsymbol{X'V_y^{-1}y}$$

- $\widehat{\boldsymbol{B}}$の分散$\boldsymbol{V_{\widehat{B}}}$

$$\boldsymbol{V_{\widehat{B}}} = \left(\boldsymbol{X'V_y^{-1}X}\right)^{-1}$$

# マルチレベルモデルの最尤推定

- ランダム効果対数$\boldsymbol{b}$の尤度関数$\log(\ell_{\boldsymbol{b}})$は次式のようになる

$$\begin{aligned}&\ell_{\boldsymbol{b}}\\&\propto \left|\left(\sigma_y^2 I\right)\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{y - XB - Zb})'\left(\sigma_y^2 I\right)^{-1}(\boldsymbol{y - XB - Zb})\right]\\&\times\left|\left(\sigma_{\boldsymbol{b}}^2 I\right)\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\boldsymbol{b'}\left(\sigma_{\boldsymbol{b}}^2 I\right)^{-1}\boldsymbol{b}\right]\end{aligned}$$

- このとき対数尤度関数$\log(\ell_{\boldsymbol{b}})$は、

$$\begin{aligned}&\log(\ell_{\boldsymbol{b}})\\&= -\frac{\log(2\pi)}{2}\\&-\frac{1}{2}\left[\log\left|\sigma_y^2 I\right| + (\boldsymbol{y - XB - Zb})'\left(\sigma_y^2 I\right)^{-1}(\boldsymbol{y - XB - Zb}) + \log\left|\sigma_{\boldsymbol{b}}^2 I\right|\right.\\&\left.+\boldsymbol{b'}\left(\sigma_{\boldsymbol{b}}^2 I\right)^{-1}\boldsymbol{b}\right]\end{aligned}$$

# マルチレベルモデルの最尤推定

- 対数尤度関数$\log(\ell_{\boldsymbol{b}})$が最大になるとき、 $\delta\log(\ell_{\boldsymbol{b}})/\delta\boldsymbol{b} = 0$

$$\begin{aligned}\delta\log(\ell_{\boldsymbol{b}})/\delta\boldsymbol{b} &= \boldsymbol{Z}'\left(\sigma_y^2 I\right)^{-1}(\boldsymbol{y} - \boldsymbol{XB} - \boldsymbol{Zb}) - \left(\sigma_{\boldsymbol{b}}^2 I\right)^{-1}\boldsymbol{b} \\ &= \boldsymbol{Z}'\left(\sigma_y^2 I\right)^{-1}(\boldsymbol{y} - \boldsymbol{XB}) - \left(\boldsymbol{Z}'\left(\sigma_y^2 I\right)^{-1}\boldsymbol{Z} + \left(\sigma_{\boldsymbol{b}}^2 I\right)^{-1}\right)\boldsymbol{b} = 0\end{aligned}$$

- ランダム効果$\boldsymbol{b}$の最尤推定量$\widehat{\boldsymbol{b}}$は、

$$\widehat{\boldsymbol{b}} = \left(\boldsymbol{Z}'\left(\sigma_y^2 I\right)^{-1}\boldsymbol{Z} + \left(\sigma_{\boldsymbol{b}}^2 I\right)^{-1}\right)^{-\mathbf{1}}\boldsymbol{Z}'\left(\sigma_y^2 I\right)^{-1}(\boldsymbol{y} - \boldsymbol{XB})$$

# マルチレベル・モデルのベイズ推定

$$y_i \sim N\left(X_i B, \sigma_y^2\right)$$

$$\begin{pmatrix}\beta_{j0} \\ \beta_{j1}\end{pmatrix} \sim N\left(\begin{pmatrix}\gamma_{00} + \gamma_{10}u_{j0} \\ \gamma_{01} + \gamma_{11}u_{j0}\end{pmatrix}, \begin{pmatrix}\sigma_{\beta_0}^2 & \rho\sigma_{\beta_0}\sigma_{\beta_1} \\ \rho\sigma_{\beta_0}\sigma_{\beta_1} & \sigma_{\beta_1}^2\end{pmatrix}\right)$$

$$B \sim N(0, V_B)$$
$$\varepsilon_{ij} \sim N\left(0, \sigma_y^2\right)$$
$$\eta_{jk} \sim N\left(0, \sigma_\eta^2\right)$$

## マルチレベル・モデルの階層ベイズ推定

- 未知パラメータは$B$を事前情報とする階層ベイズ法を用いて推定できる
- このとき$\sigma_y^2$と$V_B$の事前情報を与える必要がある
- $\sigma_y^2$の事前情報は、スケールパラメータ$s_{y_i}^2\left(=\sigma_{y_i}^2\right)$を用いて、$\chi^2$分布の逆分布として与えることができる

$$\sigma_y^2|s_{y_i}^2 \sim \frac{\eta_j s_{y_i}^2}{\chi_{\eta_j}^2}$$

- $V_B$の事前分布は、逆ウィシャート分布（後述）により与えることができる

$$V_B|\upsilon, V \sim IW(\upsilon, V)$$

- ここで、$\upsilon, V$は$V_B$の自由度と共分散のパラメータである

## マルチレベル・モデルの階層ベイズ推定

- したがって、マルチレベルモデルを階層ベイズ推定する場合は、以下のような条件付き分布を次々に生成させれば良い

$$\sigma_y^2|s_{y_i}^2 \sim \frac{\eta_j s_{y_i}^2}{\chi_{\eta_j}^2}$$
$$V_B|\upsilon, V \sim IW(\upsilon, V)$$
$$y_i|X_i, B, \sigma_y^2 \sim N\left(X_i B, \sigma_y^2\right)$$
$$B|Z_i, \sigma_{b_i}^2, V_B \sim N(0, V_B)$$
$$\boldsymbol{b_i}|\sigma_{b_i}^2 \sim N\left(0, \sigma_{b_i}^2\right)$$
$$\sigma_{b_i}^2|\eta_j \sim N\left(0, \sigma_\eta^2\right)$$

# ウィシャート分布

- $X_i$が$n$個の$p$次元多変量正規分布に従うとき

$$X_i \sim N(0, \Sigma)$$

- 標本分散共分散行列$W$は逆ウィシャート分布に従う

$$X = \sum_{i=i}^{n} X_i \cdot X_i^T$$

$$X \sim W(X; \Sigma, n, p)$$

- ベイズ統計では多変量正規分布の共分散行列の自然共役事前分布として用いられる

# ウィシャート分布

- ウィシャート行列はカイ二乗分布を多次元化したものである

$$W(X; \Sigma, n, p) = \frac{|\Sigma|^{-\frac{n}{2}} |X|^{\frac{n-p-1}{2}}}{2^{\frac{np}{2}} \cdot \Gamma_p\left(\frac{n}{2}\right)} \cdot exp\left(-\frac{1}{2} tr\Sigma^{-1} X\right)$$

- $\Gamma_p(a)$は$p$次元多変量Γ関数

$$\Gamma_p(a) = \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \left[a - \frac{1}{2}(j-1)\right]$$

# 逆ウィシャート分布

- $W(X;\Sigma,n,p)$に従う$X$を$\mathrm{V}=X^{-1}$とすると、Vの分布を逆ウィシャート分布という

$$W(V;\Sigma,n,p)=\frac{|\Sigma|^{-\frac{n}{2}}|V|^{-\frac{n-p-1}{2}}}{2^{\frac{np}{2}}\cdot\Gamma_p\left(\frac{n}{2}\right)}\cdot exp\left(-\frac{1}{2}tr\Sigma^{-1}V^{-1}\right)$$