多変量解析法の分類：　基準変数 $\boldsymbol{Y}$ がカテゴリカルな場合

		説明変数 $\boldsymbol{X}$は観測されているか	
		観測されている （顕在変数）	観測されていない （潜在変数）
説明変数 $\boldsymbol{X}$ は連続か離散か	連続	ロジスティク回帰分析 Multimomial Logit, Proportional Odds, Probit Model Logistic Discriminant Analysis	カテゴリカルデータの因子分析・主成分分析 IRT { 数量化 III 類 }
	離散	対数線形モデル	潜在クラス分析

データとそのダミー変数への展開

$$y_{ij} = \{A_0, A_1, \ldots, A_{m_j}\} \tag{4}$$

$$z_{ijk} = \begin{cases} 1, & \text{if} \quad y_{ij} = A_k \\ 0, & \text{if} \quad y_{ij} \neq A_k \end{cases}, k = 0, 1, \ldots, m_j \tag{5}$$

モデル

$$\begin{aligned} \Pr\left(y_{ij} = A_k | \boldsymbol{x}_i\right) &= \Pr\left(z_{ijk} = 1 | \boldsymbol{x}_i\right) \\ \boldsymbol{\mathcal{E}}(z_{ijk} | \boldsymbol{x}_i) &= \pi_{ijk}, \quad \sum_{k=0}^{m_j} \pi_{ijk} = 1 \end{aligned}$$

$$\pi_{ijk} = \frac{f(\boldsymbol{x}_i | \boldsymbol{\beta}_{jk})}{\sum_{k=0}^{m_j} f(\boldsymbol{x}_i | \boldsymbol{\beta}_{jk})}, \quad f > 0 \tag{6}$$

尤度関数は多項分布の積：最尤法

$$\boldsymbol{z}_{ij} \sim MN(\boldsymbol{\pi}_{ij}, 1), \quad L = \prod_{i=1}^{N} \prod_{j=1}^{p} \prod_{k=0}^{m_j} \pi_{ijk}^{z_{ijk}} \tag{7}$$

ロジットとロジスティック関数は互いに逆関数

Logit Transformation

$$z = \log\left(\frac{\pi}{1-\pi}\right), \quad 0 < \pi < 1$$

Logistic Transformation

$$\pi = \frac{\exp(z)}{(1+\exp(z))} = \frac{1}{(1+\exp(-z))}$$

ロジスティク回帰モデル（２値項目反応理論）

$$f(\boldsymbol{x_i}|\boldsymbol{\beta}_{jk}) = \exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{jk}) \tag{8}$$

とすると、これは π_{ijk} の logit に線形モデル

$$logit(\pi_{ijk}) \equiv ln\frac{\pi_{ijk}}{1-\pi_{ijk}} = \boldsymbol{x_i}'\boldsymbol{\beta}_{jk} \tag{9}$$

を考えていることとなる。（特に、 $p = 1, m_1 = 1$ の場合）

$$\pi_{ij1} = \frac{\exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{j1})}{\exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{j0}) + \exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{j1})} \tag{10}$$

分子分母に定数を掛けても代わらない（乗算的不定性）

$$\beta_{j0} = 0 \tag{11}$$

$$\pi_{ij1} = \frac{\exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{j1})}{\exp(0) + \exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{j1})} \tag{12}$$

$$= \frac{\exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{j1})}{1+\exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{j1})} \tag{13}$$

$$= \frac{1}{1+\exp(-\boldsymbol{x_i}'\boldsymbol{\beta}_{j1})} \tag{14}$$

Binary Logistic Model

$$P_j(\theta) = c_j + (1-c_j) \times \frac{1}{1+\exp(-1.7a_j(\theta - b_j))} \tag{15}$$

Multinomial Logit Model（名義尺度項目反応理論）

$$f(\boldsymbol{x_i}|\boldsymbol{\beta}_{jk}) \quad = \quad \exp(\boldsymbol{x_i}'\boldsymbol{\beta}_{jk}) \tag{16}$$

とすると、これは π_{ijk} の logit に線形モデル

$$logit(\pi_{ijk}) \equiv ln\frac{\pi_{ijk}}{\pi_{ijm_j}} = \boldsymbol{x_i}'\boldsymbol{\beta}_{jk} \tag{17}$$

を考えていることとなる。

分子分母に定数を掛けても代わらない（乗算的不定性）

$$\beta_{j0} = 0 \qquad \text{or} \qquad \beta_{jm_j} = 0 \tag{18}$$

Proportional Odds Model（段階反応モデル）

$$\begin{aligned}
\Pr(y_j = A_k|\boldsymbol{x}) \quad &= \quad \pi_{jk} \;=\; P^+_{jk}(\boldsymbol{x}) - P^+_{jk+1}(\boldsymbol{x}), \\
&\qquad k = 0, 1, 2, \ldots, m_j \\
P^+_{j0}(\boldsymbol{x}) \quad &= \quad 1 \qquad \text{and} \qquad P^+_{jm_j+1}(\boldsymbol{x}) \;=\; 0 \\
P^+_{jk}(\boldsymbol{x}) \quad &= \quad \frac{1}{1+\exp(-\boldsymbol{x}\boldsymbol{\beta}_{jk}))}, \\
&\qquad k = 1, 2, \ldots, m_j
\end{aligned}$$

Graded Response Model

$$\begin{aligned}
\Pr(U_j = k|\theta) \qquad &= \qquad P_{jk}(\theta) \;=\; P^+_{jk}(\theta) - P^+_{jk+1}(\theta), \\
&\qquad\qquad k = 0, 1, 2, \ldots, m_j \\
P^+_{j0}(\theta) \qquad &= \qquad 1 \qquad \text{and} \qquad P^+_{jm_j+1}(\theta) \;=\; 0 \\
P^+_{jk}(\theta) \qquad &= \qquad \frac{1}{1+\exp(-1.7a_j(\theta - b_{jk}))}, \\
&\qquad\qquad k = 1, 2, \ldots, m_j \\
&\qquad \text{または} \\
P^+_{jk}(\theta) \qquad &= \qquad \frac{1}{1+\exp(-1.7a_j(\theta - b_j + c_{jk}))}, \\
&\qquad\qquad k = 1, 2, \ldots, m_j, \sum_{k=1}^{m_j} c_{jk} = 0
\end{aligned}$$

1　項目反応理論とプロビット回帰

1.1　2 値データ

測ろうとする１次元の特性を考える。そして、その上に各被験者がその特性値の大きさ（これを θ と記す）により位置づけられているとする。また、各項目（$j = 1, 2, \ldots, p$) もその難易度（これを b_j とする）に応じてその上に並んでいるとする。また、項目には、精度の善し悪しがあるとし、その精度の悪さを ψ_j で表すとする。そして、直接的に b_j と θ を比較するのではなく、θ に項目の精度を反映した誤差が加わったもの

$$Y_j = \theta + e_j, \quad e_j \sim N(0, \psi_j^2) \tag{19}$$

$$Y_j|\theta \sim N(\theta, \psi_j^2) \tag{20}$$

と b_j の値が比較されるものとする。

$$Y_j \geq b_j \quad \text{ならば} \quad U_j = 1 \quad （項目 j に正答） \tag{21}$$

$$Y_j < b_j \quad \text{ならば} \quad U_j = 0 \quad （項目 j に誤答） \tag{22}$$

となる。これを図示したものが Figure 1 である。この場合、誤差の変動幅 ψ_j が大きければ

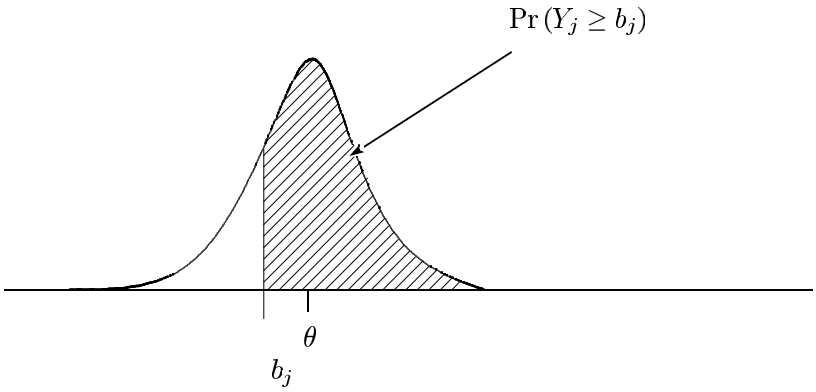

Figure 1: Thurstone Model of IRT

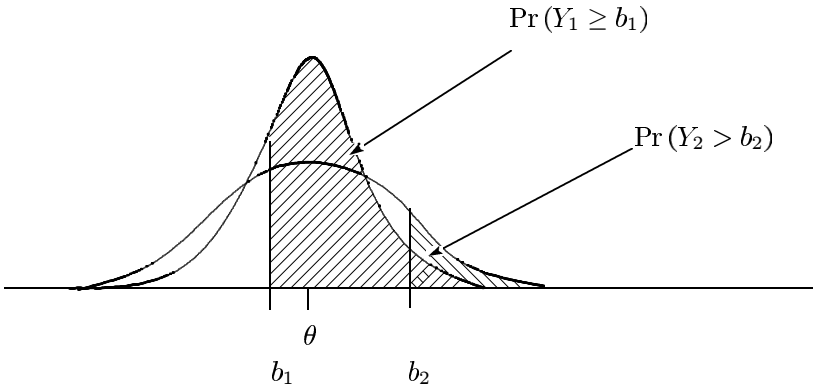

Figure 2: Thurstone Model of IRT

Y_j は θ と異なる値をとることがあるため、その大きさが、その項目の不正確さを表していることになる。これを図示したものが Figure 2 である。

このモデルによると、たとえ、本来は $\theta \leq b_j$ である場合でも、そのときの Y_j の値により、その項目に正答する場合もあり得ることになる。

上記の仮定から、特性値が θ の人が困難度 b_j の項目に正答する確率というものを計算することが出来る。

$$\begin{aligned}
\Pr(U_j = 1|\theta) &= \Pr(Y_j \geq b_j) \\
&= \Pr\left(\tfrac{Y_j-\theta}{\psi_j} > \tfrac{b_j-\theta}{\psi_j}\right) \\
& \qquad\qquad Y_j \sim N(\theta, \psi_j^2) \\
& \qquad\qquad Z_j = \tfrac{Y-\theta}{\psi_j} \sim N(0,1) \\
&= \Pr\left(Z_j > \tfrac{b_j-\theta}{\psi_j}\right) \\
& \qquad\qquad \text{対称性} \\
&= \Pr\left(Z_j < -\tfrac{b_j-\theta}{\psi_j}\right) \\
&= \Phi\left(\tfrac{\theta-b_j}{\psi_j}\right)
\end{aligned} \tag{23}$$

となる。

しかし、通常は、項目の精度を表すために $a_j = 1/\psi_j$ を用い、累積正規分布関数をロジスティク曲線で近似したもの

$$\Pr(U_j = 1|\theta) = P_j(\theta) = \frac{1}{1+\exp(-1.7a_j(\theta - b_j))} \tag{24}$$

を用いることが多い。この $P_j(\theta)$ は2パラメタロジスティク項目特性曲線 (Item Characteristics Curve) と呼ばれているが、これは、特性値が θ の人が項目 j に正答する確率を与える関数である。（特性値 θ を持つ人が、その項目に答えたということを忘れながら、何度もその項目に答えた場合の平均正答率に当たる。）また、a_j を項目 j の識別力パラメタ (discrimination/slope)、b_j を困難度パラメタ (difficulty/threshold) と呼ぶ。

1.2 カテゴリカルデータの因子分析、プロビット回帰

1 因子モデル

$$Y_j = \mu_j + \lambda_j\theta + e_j, \quad e_j \sim N(0, \psi_j^2) \tag{25}$$

$$Y_j|\theta \sim N(\mu_j + \lambda_j\theta, \psi_j^2) \tag{26}$$

を考える。カテゴリカルデータ U_j の発生は

$$Y_j \geq \gamma_j \quad \text{ならば} \quad U_j = 1 \tag{27}$$

$$Y_j < \gamma_j \quad \text{ならば} \quad U_j = 0 \tag{28}$$

であるとする。これを図示したものが Figure 4 である。

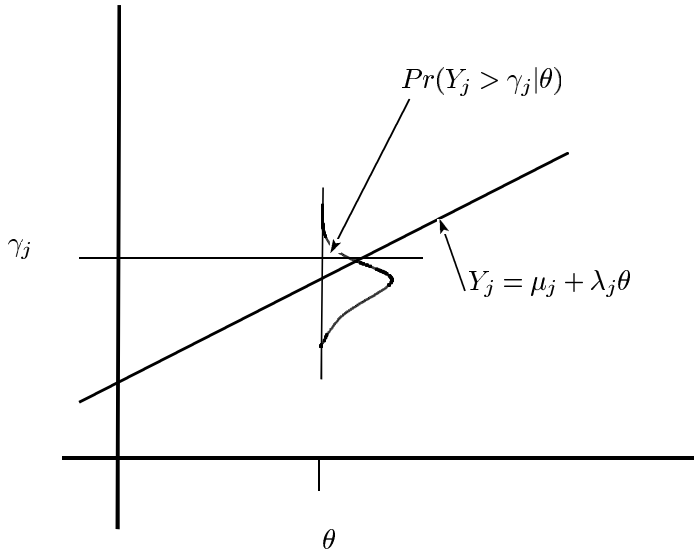

Figure 3: Categorical FA Model

このとき正反応データの発生確率は

$$\begin{aligned}
\Pr(U_j = 1|\theta) &= \Pr(Y_j \geq \gamma_j) \\
&= \Pr\left(\tfrac{Y_j-\mu_j-\lambda_j\theta}{\psi_j} > \tfrac{\gamma_j-\mu_j-\lambda_j\theta}{\psi_j}\right) \\
& \qquad\qquad Y_j \sim N(\mu_j - \lambda_j\theta, \psi_j^2) \\
& \qquad\qquad Z_j = \tfrac{Y-\mu_j-\lambda_j\theta}{\psi_j} \sim N(0,1) \\
&= \Pr\left(Z_j > \tfrac{\gamma_j-\mu_j-\lambda_j\theta}{\psi_j}\right) \qquad (29) \\
& \qquad\qquad \text{対称性} \\
&= \Pr\left(Z_j < -\tfrac{\gamma_j-\mu_j-\lambda_j\theta}{\psi_j}\right) \\
&= \Phi\left(\tfrac{\lambda_j\theta-(\gamma_j-\mu_j)}{\psi_j}\right)
\end{aligned}$$

となるが、パラメタは識別されない。

$$a_j = \frac{\lambda_j}{\psi_j} \qquad (30)$$

$$b_j = \frac{(\gamma_j - \mu_j)}{\psi_j} \qquad (31)$$

を用いれば、項目反応理論の形となる。

多値データの場合を図示したものが Figure 4 である。

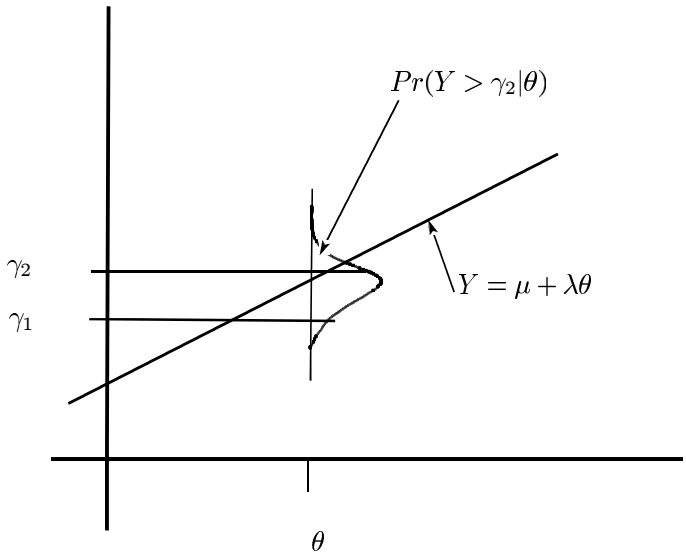

Figure 4: Graded Response Model