平成 23 年 06 月 16 日
兵藤 哲朗

Ordered Logit/ Ordered Probit モデルとその推定

■Ordered とは？

顧客満足度データや，車の保有台数など，その数字に順序だった関係がある場合，それら変数を，Ordered（順序）変数という．Ordered 変数を被説明変数とした回帰モデルを推定することも考えられるが，暗に，『満足度=3 は，満足度=1 の 3 倍』など，根拠のない扱いを余儀なくされる．これを回避する一手法が，Ordered Logit Model，もしくは，Ordered Probit Model であり，通常の MNL や MNP の効用項の解釈を工夫するだけで容易にモデル構築を行うことができる．

■潜在変数の仮定

いま，説明変数と，未知パラメータで構成される潜在変数（Latent variable）を考えよう．これは MNL と同様，無次元の効用項とみなしてもよい．満足度アンケートの回答結果を説明するモデルを仮定し，その選択肢数は 3 であったとする（「3:満足，2:どちらとも言えない，1:不満」など）．潜在変数を，

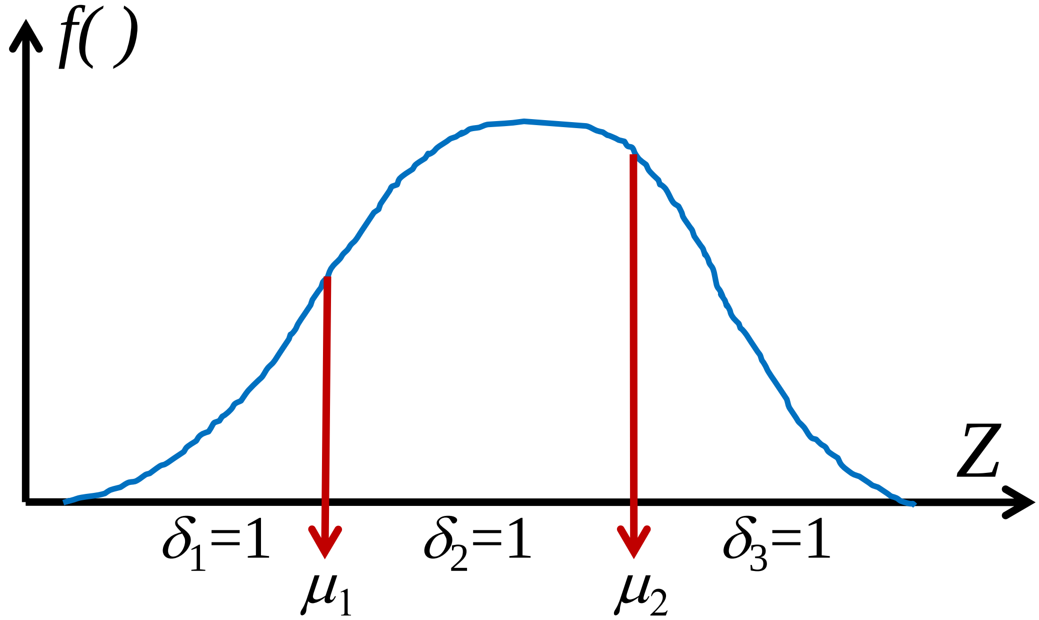
$$y_n^* = \sum_k \theta_k x_{nk} + \varepsilon_n = Z_n + \varepsilon_n \tag{1}$$

とし，選択結果を以下で表現する．

$$\delta_{n1} = 1,\ \text{if}\ \ -\infty \le y_n^* \le \mu_1 \tag{2}$$

$$\delta_{n2} = 1,\ \text{if}\ \ \mu_1 \le y_n^* \le \mu_2 \tag{3}$$

$$\delta_{n3} = 1,\ \text{if}\ \ \mu_2 \le y_n^* \le \infty \tag{4}$$



上の図から分かるとおり，潜在変数は確率的な誤差項を含むため，確率分布しており，それ故，潜在変数が各選択肢の定義区間に含まれる確率を，選択確率と見なせることになる．

ここで ε に正規分布を仮定すれば，Probit モデルと同様に，Ordered Probit モデルとなる．また，Gumbel 分布を仮定すれば Logit モデルを用いた式となる．標準正規分布の分布関数を $F(\cdot)$ とすれば，これらは

以下のように整理できる（個人の添え字 n は省略）.

Ordered Logit

$$P_1 = \frac{1}{1+\exp[Z-\mu_1]}$$

$$P_2 = \frac{1}{1+\exp[Z-\mu_2]} - \frac{1}{1+\exp[Z-\mu_1]}$$

$$P_3 = 1 - \frac{1}{1+\exp[Z-\mu_2]}$$

Ordered Probit

$$P_1 = F(\mu_1 - Z)$$

$$P_2 = F(\mu_2 - Z) - F(\mu_1 - Z)$$

$$P_3 = 1 - F(\mu_2 - Z)$$

あとは通常の MNL と同様，以下の尤度関数を最大化すればよい.

$$L = \sum_n \sum_i \delta ni \cdot \ln(P_{ni}) \tag{5}$$

ちなみに，Ordered Logit モデルも，Ordered Probit モデルも，対数尤度関数のパラメータに関するヘッセ行列は正定値であることが証明でき，単峰性があるため，唯一解を容易に推定可能である．ただし，区切りパラメータについては，$\mu_i \le \mu_{i+1}$ という大小関係の制約条件がある．一般に，最大化問題を解く課程では，勾配最大方向に一定の距離までパラメータ値を伸ばし，黄金分割アルゴリズムなどを用いて，ステップ間の最大値を求める．その値を伸ばす時に，上記の制約が満たされないことがあり，パラメータ推定途中でエラーが出る可能性もある．これを回避するためには，

$$\mu_{i+1} = \mu_i + \exp(\alpha_{i+1}) \tag{6}$$

として，α を未知パラメータとして推定すればよい.

さて，具体的なデータを用いた事例を紹介しよう．用いたのは，本年 5 月に，流通情報工学科 3 年生に対して行ったアンケートである.

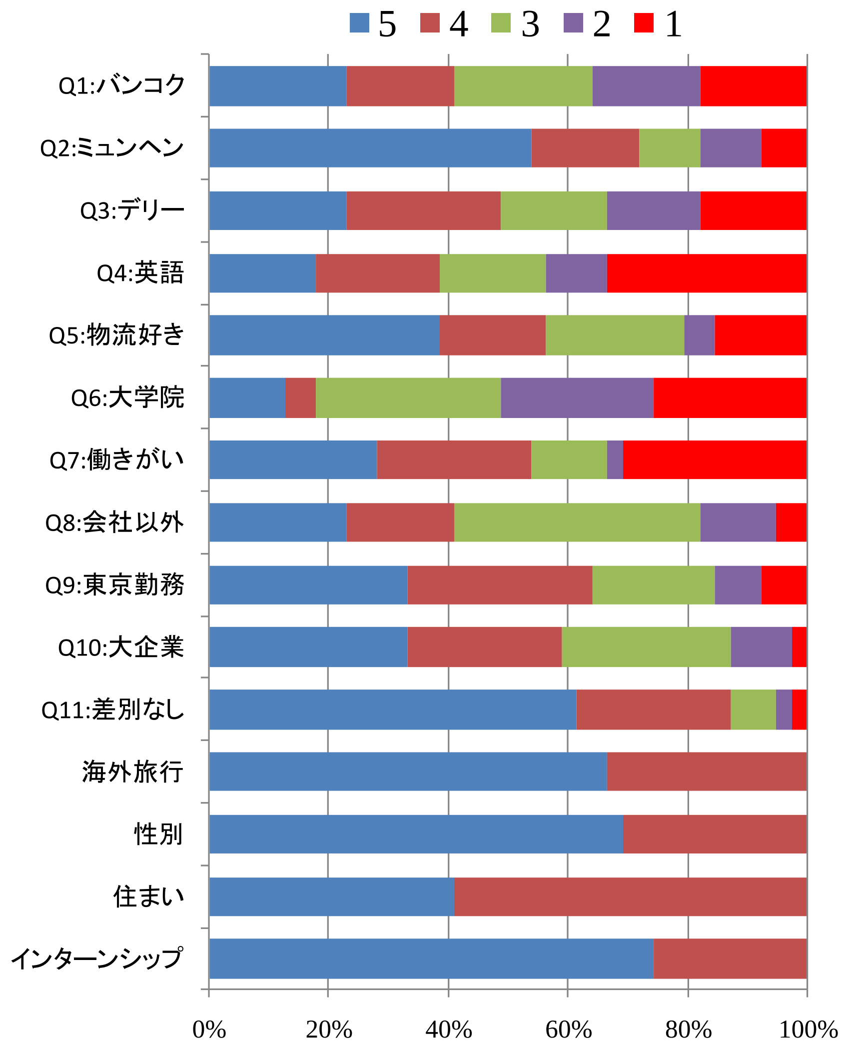
最近の若者は内向きになり，あまり海外にも行きたがらないといわれる．しかしながら，物流やロジスティクス業では，アジアをはじめとする諸外国が主活動領域になりつつあり，内向きではとても生き抜くことはできない．それはともかく，アンケートでは，バンコク・ミュンヘン・デリーへの海外赴任を受け入れるか否か　という問いがあり，それについて 5 段階評価が回答される.

アンケート票と，その単純集計は次ページに掲げるが，ミュンヘンへの赴任意向高いのはともかく，バンコクとデリーでは差がないのが面白い．あまりバンコクの便利さも，デリーの衛生状況なども知られていないようである.

就職観＆海外赴任意向アンケート

	そう思う	どちらかといえばそう思う	どちらともいえない	どちらかといえば思わない	そう思わない
Q1：入社５年後，タイ・バンコクへの海外赴任を受け入れる	5	4	3	2	1
Q2：入社５年後，ドイツ・ミュンヘンへの海外赴任を受け入れる	5	4	3	2	1
Q3：入社５年後，インド・デリーへの海外赴任を受け入れる	5	4	3	2	1
Q4：英語は比較的得意，または英語の勉強は苦にならない	5	4	3	2	1
Q5：情報工学系の授業より，物流系の授業が得意だし好きだ	5	4	3	2	1
Q6：大学院に進学したい	5	4	3	2	1
Q7：多少収入額が高い仕事より，『働きがいのある』仕事に就きたい	5	4	3	2	1
Q8：アフターファイブは会社以外の人間とつきあいたい	5	4	3	2	1
Q9：ずっと東京近辺で働きたい	5	4	3	2	1
Q10：できれば大企業に勤めたい	5	4	3	2	1
Q11：大学・男女差別のない会社に勤めたい	5	4	3	2	1

海外旅行の経験はありますか？	5.有り	4.なし
性別は？	5.男性	4.女性
今のお住まいは？	5.自宅から	4.アパート，寮
夏のインターンシップは？	5.行きたい	4.行かない



回答総数：39 サンプル

このアンケート結果のテキストデータを以下に掲げる．見ての通り，csv ファイルである．

```
,Bangkok,Munich,Delhi,English,Logistics,Gratuate,Kai,Igai,To
kyo,Big,Sabetsu,Abroad,Sex,House,Intern
1,3,5,3,1,1,5,2,2,4,5,5,0,1,0,0
2,3,4,4,3,3,1,4,3,5,3,5,0,1,0,1
3,1,2,1,3,4,1,3,3,4,5,5,1,1,1,1
4,3,5,4,2,1,2,3,2,4,4,5,0,1,1,0
5,2,3,2,1,5,2,4,4,3,3,4,1,1,0,1
6,2,3,3,4,5,3,2,4,4,5,4,1,1,0,1
7,2,4,1,1,4,5,5,5,5,3,5,1,0,0,1
8,5,5,5,5,4,3,4,4,3,3,5,1,1,1,1
9,1,5,1,1,3,5,1,5,5,5,3,1,1,0,0
10,4,5,4,5,5,1,2,3,5,3,5,0,1,1,1
11,5,5,5,1,3,3,5,3,3,5,5,1,1,0,1
12,1,5,2,5,5,1,4,1,1,2,5,0,1,1,0
13,2,2,2,1,2,3,4,4,4,4,5,0,1,1,1
14,3,5,3,4,5,3,5,3,3,2,4,0,1,0,0
15,2,5,2,4,1,3,4,3,5,5,5,0,1,0,1
16,3,3,3,3,3,4,3,3,4,5,3,1,1,0,1
17,5,1,5,4,3,3,5,5,1,1,5,1,1,0,1
18,4,4,4,1,1,3,5,1,5,4,5,1,1,0,0
19,2,2,2,1,3,2,4,3,4,4,4,1,1,0,0
20,4,5,4,4,5,3,5,4,4,4,4,1,1,1,1
21,1,2,1,1,1,2,5,3,4,3,5,1,1,0,1
22,1,4,4,3,4,5,5,4,3,4,5,1,1,1,0
23,5,5,5,4,5,3,5,3,3,4,5,0,1,0,1
24,3,3,3,3,3,2,4,5,4,5,3,0,1,1,0
25,5,5,5,5,5,2,5,3,3,5,5,1,1,1,1
26,4,5,4,2,5,4,5,2,5,4,5,1,0,0,1
27,5,5,5,2,5,1,3,2,5,5,4,1,1,0,1
28,4,5,4,1,5,3,3,5,3,3,1,1,1,0,0
29,2,4,1,3,3,2,2,5,5,4,4,1,0,1,1
30,1,1,1,5,3,1,3,5,5,3,5,1,1,1,1
31,3,5,2,1,4,1,3,3,2,3,4,1,0,1,1
32,4,5,4,4,4,2,4,3,4,4,4,1,0,0,1
33,5,5,5,3,5,1,3,3,2,5,5,1,0,0,1
34,4,4,3,2,2,2,5,3,5,2,5,1,0,1,1
35,1,1,1,1,1,3,3,4,5,3,5,0,0,1,1
36,3,5,4,4,5,2,3,2,2,3,4,1,0,0,1
37,3,4,3,1,5,1,2,3,4,2,2,0,0,0,1
38,5,5,5,5,4,1,4,5,5,5,5,1,0,1,1
39,5,5,5,5,5,5,4,5,1,5,5,0,0,0,1
```

さて，無理矢理，赴任意向（1～5 の値）を被説明変数として重回帰分析を行うと，以下の通りとなった．

	バンコク		ミュンヘン		デリー	
	回帰統計		回帰統計		回帰統計	
	重相関 R	0.5878	重相関 R	0.6989	重相関 R	0.6423
	重決定 R2	0.3455	重決定 R	0.4885	重決定 R	0.4126
	補正 R2	0.0434	補正 R2	0.2525	補正 R2	0.1415
	標準誤差	1.3976	標準誤差	1.1566	標準誤差	1.3306
	観測数	39	観測数	39	観測数	39
	係数	t	係数	t	係数	t
	1.963	0.832	3.782	1.937	1.656	0.737
Q4:英語	0.207	1.018	0.190	1.129	0.288	1.490
Q5:物流好き	0.077	0.342	0.355	1.913	0.154	0.720
Q6:大学院	-0.095	-0.412	0.225	1.171	-0.023	-0.104
Q7:働きがい	0.470	1.902	0.081	0.396	0.412	1.752
Q8:会社以外	-0.134	-0.528	-0.521	-2.486	-0.227	-0.944
Q9:東京勤務	-0.155	-0.706	0.082	0.453	-0.209	-0.998
Q10:大企業	0.337	1.455	0.307	1.604	0.344	1.564
Q11:差別なし	-0.288	-0.774	-0.215	-0.697	-0.168	-0.475
海外旅行	-0.098	-0.187	-0.282	-0.651	-0.099	-0.198
性別	-0.475	-0.858	-0.873	-1.906	-0.143	-0.271
住まい	-0.458	-0.804	0.087	0.185	-0.417	-0.770
インターンシップ	0.538	0.770	-0.599	-1.036	0.073	0.109

自由度調整済みの決定係数をみても，あまり説明力のないモデルである．これは Ordered モデルにしても同様であろう．「英語好き」は，ほどほど共通して t 値は高めである．ミュンヘンは特に女性に人気のようだ．

このデータに対する，R の Ordered Logit モデル推定プログラムは次ページの通り．

```
dt <- read.csv("c:/usr/doc/kaigai.csv",header=TRUE)
hh <- nrow(dt)

tgt <- 3 #被説明変数の変数番号
alt <- 5 #被説明変数のカテゴリーの数
np  <- 5 #説明変数（Z）の変数の数
b0  <- array(0, (np+alt-1) )
Cshr <- array(0,alt); L0 <- 0
for (i in 1:alt) {
  Cshr[i] <- sum(dt[,tgt]==i); cat("    #",i,": ",Cshr[i]); L0 <- L0 +
log(Cshr[i]/hh)*Cshr[i]
}; cat("¥n")

fr <- function(x) {
LL=0
Z <- x[1]*dt[,5] + x[2]*dt[,6] + x[3]*dt[,11] + x[4]*dt[,13] + x[5]*dt[,14]
b <- array(0, alt-1); b[1] <- x[np+1]
for (i in 2:(alt-1)) b[i] <- b[i-1] + exp(x[np+i])
p <- array(0,c(hh,alt))
p[,1] <- 1/( 1+exp(Z-b[1]) )
for (j in 2:(alt-1)) {
  p[,j] <- 1/( 1+exp( Z-b[j] ) ) - 1/( 1+exp( Z-b[j-1] ) )
}
p[,alt] <- 1 - 1/( 1+exp( Z-b[alt-1] ) )
pp <- array(0,c(hh,alt)); dc <- array(0,c(hh,alt))
for (j in 1:alt) {
  pp[,j] <- (p[,j]!=0)*p[,j] + (p[,j]==0)
  dc[,j] <- dt[,tgt]==j
}

LL <- sum( dc*log(pp) )
return(LL)
}

fr2 <- function(x) {
LL=0
Z <- x[1]*dt[,5] + x[2]*dt[,6] + x[3]*dt[,11] + x[4]*dt[,13] + x[5]*dt[,14]
b <- x[(np+1):(np+alt-1)]
p <- array(0,c(hh,alt))
p[,1] <- 1/( 1+exp(Z-b[1]) )
for (j in 2:(alt-1)) {
  p[,j] <- 1/( 1+exp( Z-b[j] ) ) - 1/( 1+exp( Z-b[j-1] ) )
}
p[,alt] <- 1 - 1/( 1+exp( Z-b[alt-1] ) )
pp <- array(0,c(hh,alt)); dc <- array(0,c(hh,alt))
for (j in 1:alt) {
  pp[,j] <- (p[,j]!=0)*p[,j] + (p[,j]==0)
  dc[,j] <- dt[,tgt]==j
}
LL <- sum( dc*log(pp) )
return(LL)
}

res<-optim(b0,fr, method = "BFGS", hessian = TRUE, control=list(fnscale=-1))
print(res$par); b0 <- res$par
for (i in (np+2):(np+alt-1)) b0[i] <- b0[i-1] + exp(b0[i])

res<-optim(b0,fr2, method = "BFGS", hessian = TRUE, control=list(fnscale=-1))
b   <- res$par; hhh  <- res$hessian; tval <- b/sqrt(-diag(solve(hhh)))

print(res$par); print(tval); LL <- res$value
cat(" roh =   ",(L0-LL)/L0,"¥n")
cat(" rohbar= ",(L0-(LL-length(b)))/L0,"¥n")
```

プログラムには，二つの function 文（対数尤度の算出関数）が定義されている．最初の function 文では，区切りパラメータ大小関係の制約を満たすため，(6)式を用いて，α を推定する．その後，t 値計算のためのヘッセ行列推定を，$\mu_i = \exp(\alpha_i)$としてパラメータを再推定しているのである．ちょっと無骨な方法だ.

このプログラムを用いて，回帰分析の結果も参考にしながら，以下の 5 説明変数と 4 つの区切りパラメータの，合計 9 変数を推定した.

バンコク：(パラメータ下段は t 値)

Q4.英語	Q5.物流	Q10.大企業	海外旅行	性別	μ_1	μ_2	μ_3	μ_4
0.2375	0.3539	0.3553	-0.0254	-0.5356	1.1694	2.1838	3.2487	4.2495
1.06	1.43	1.11	-0.038	-0.82	0.74	1.38	2.01	2.51

roh = 0.0600
rohbar= -0.0840

ミュンヘン：(パラメータ下段は t 値)

Q4.英語	Q5.物流	Q10.大企業	海外旅行	性別	μ_1	μ_2	μ_3	μ_4
0.1377	0.5432	0.6699	-0.9024	-0.8971	0.5905	1.7883	2.4895	3.4303
0.55	2.07	2.01	-1.19	-1.15	0.35	1.07	1.50	2.04

roh = 0.1164272
rohbar= -0.06026387

デリー：(パラメータ下段は t 値)

Q4.英語	Q5.物流	Q10.大企業	海外旅行	性別	μ_1	μ_2	μ_3	μ_4
0.3501	0.4029	0.3874	-0.0429	-0.0205	2.0380	2.9821	3.8703	5.2748
1.53	1.62	1.25	-0.065	-0.031	1.327	1.923	2.43	3.11

roh = 0.0840623
rohbar= -0.06090088

尤度比を見ても分かるとおり，説明力は小さく，不十分なモデルである．しかし，区切り変数の大小関係は守られていることが分かる.

ここでは Ordered Logit のみの具体例となったが, Ordered Probit モデルでは, 標準正規分布関数である,

```
pnorm(x[1]-Z, mean=0,sd=1)
```

を Ordered Logit の相当部分と置き換えるだけで容易にモデル推定は可能であろう.

以上