R による多項ロジット、プロビットモデルの推定

多項ロジット(multinomial logit)、プロビット(probit)モデルを R で推定する方法について説明する。多項ロジットモデルについては、別の資料を参照のこと。

1．パッケージ mlogit のインスツール

多項ロジットモデルを推定するために R のパッケージ mlogit をあらかじめインスツールしなければならない。パッケージとは通常の R には含まれていない、追加的な R のコマンドの集まりのようなものである。R には追加的に 600 以上のパッケージが用意されており、それぞれ分析の目的に応じて標準の R にパッケージを追加していくことになる。

インターネットに接続してあるパソコンで R を起動させ、コマンドウィンドウ(R Console)で

```
> options(CRAN="http://cran.r-project.org")
> install.packages("mlogit")
```

と入力する。すると(以下の部分は人によっては表示されないかもしれない)

「パッケージをインスツールするために個人的なライブラリ
'C:¥Users¥ …'
を作りたいですか？」

という質問が出てくるので('C:¥Users¥ …'の部分は個人個人で異なる)「はい(Y)」をクリックする。

すると「CRAN mirror」というものが出てくるので、そこから「Japan(Tsukuba)」を選び「OK」をクリックする。すると R のコマンドウィンドウにインスツールの途中経過が表示され

…
パッケージ 'lmtest' は無事に開封され、MD5 サムもチェックされました
パッケージ 'maxLik' は無事に開封され、MD5 サムもチェックされました
パッケージ 'zoo' は無事に開封され、MD5 サムもチェックされました
パッケージ 'mlogit' は無事に開封され、MD5 サムもチェックされました

ダウンロードされたパッケージは、以下にあります
C:¥Users …

のように表示される。次にダウンロードしたパッケージを使うためにコマンドウィンドウに

```
> library("mlogit")
```

と入力すると(再びコマンドウィンドウ上にいろいろと表示され)パッケージ mlogit を使用できる様になる。

2. 分析の準備、データの読み込み

データは flabordata.txt を用いる。これはアメリカの既婚女性の就業に関するデータである(データについて詳しくは後述)。このデータを読み込む。まず、「ファイル」→「ディレクトリの変更」とクリックしていくことにより、データのおいてあるフォルダ(ディレクトリ)へ移動する。そして

```
> flabordata = read.table("flabordata.txt",header=T,skip=12)
```

によってデータを読み込む。データは 3382 人の既婚女性の就業に関するデータであり、

```
> head(flabordata, 7)
  hour h.choise income  age edu n1 n2 n3 race home lur
1 2000        5    350   26   12  0   1  0     0    1    7
2  390        2    241   29    8   0   1  1     0    1    4
3 1900        5    160   33   10  0   2   0    0    1    7
4    0        1     80   20    9  2   0   0    0    1    7
5 3177        6    456   33   12  0   2   0    0    1    7
6    0        1    390   22   12  2   0   0    0    1    7
7    0        1    181   41    9  0   0   1    0    1    7
```

のようになっている(ここで head(データの名前,X) は該当データの上から X 行を表示するコマンド)。一番左は R によってデータにつけられた番号である。hour はその既婚女性の就業時間を表してる。h.choice はその既婚女性が就業時間に関してどのような選択をしたかを表し、hour=0(つまり働いていない)を選択したのであれば h.choice=1 を、0<hour≤750 を選択したのであれば h.choice=2 を、750<hour≤1250 であれば h.choice=3 を、1250<hour≤1750 であれば h.choice=4 を、1750<hour≤2250 であれば h.choice=5 を、2250<hour であれば h.choice=6 をとる変数であるとする。income, age, edu, n1, ... ,lur はそれぞれの既婚女性の特性を表すデータである(これらの詳しい説明は flabordata.txt 内の説明を参照のこと)。

このデータは 1 行に 1 個人についてのデータが並んでいる事に注意しよう。このような形式のデータは "wide" 形式と呼ばれる(ここでは元のテータは常に "wide" 形式をとるとする)。

ここでは多項ロジットモデルによって、説明変数 income, age, edu, n1, ... ,lur が既婚女性の就業時間の選択にどのような影響を及ぼすかを分析する。まず、上記のように読み込んだデータをさらに mlogit.data というコマンドを用いて mlogit で扱うデータの形式に直す。

```
> flabor = mlogit.data(flabordata, shape = "wide", choice = "h.choice")
```

ここで flabordata は読み込んだもともとのデータの名前であり、shape = "wide" はもともとのデータが"wide" 形式である事を指定している。 choice = "h.choice" は個々の選択肢が並んでいる列を知らせる(この場合 "h.choice" の列に選択の結果が並んでいるのでこうする。一般的には choice = "XXX" で XXX のところに選択肢の列の名前が入る)。この flabor は

```
> head(flabor,13)
     hour h.choice income age edu n1 n2 n3 race home lur chid alt
1.1 2000    FALSE    350   26  12  0  1  0    0    1   7    1   1
1.2 2000    FALSE    350   26  12  0  1  0    0    1   7    1   2
1.3 2000    FALSE    350   26  12  0  1  0    0    1   7    1   3
1.4 2000    FALSE    350   26  12  0  1  0    0    1   7    1   4
1.5 2000     TRUE    350   26  12  0  1  0    0    1   7    1   5
1.6 2000    FALSE    350   26  12  0  1  0    0    1   7    1   6
2.1  390    FALSE    241   29   8  0  1  1    0    1   4    2   1
2.2  390     TRUE    241   29   8  0  1  1    0    1   4    2   2
2.3  390    FALSE    241   29   8  0  1  1    0    1   4    2   3
2.4  390    FALSE    241   29   8  0  1  1    0    1   4    2   4
2.5  390    FALSE    241   29   8  0  1  1    0    1   4    2   5
2.6  390    FALSE    241   29   8  0  1  1    0    1   4    2   6
3.1 1900    FALSE    160   33  10  0  2  0    0    1   7    3   1
```

のような形式のデータになっている。ここで一番左の列の、例えば、 1.1 はその行のデータが個人 1 の選択 1 についてのデータである事を表し、1.4 はその行のデータが個人 1 の選択 4 についてのデータを表している。2 列目の choice はその選択肢が選ばれたら TRUE, 選ばれなかったら FALSE になる。一番右の chid, alt というのは chid は個人の番号、 alt というのは選択肢を表している。

3. mlogit による多項ロジットモデルの推定(説明変数が個人 *i* のみに依存している場合)

前節で準備したデータに対して、多項ロジットモデルを推定する。この問題では説明変数は個人 *i* のみに依存しているので(説明変数が選択肢 *j* にも依存する場合は次節で説明する)以下のコマンドで推定する。

```
> result = mlogit(h.choice~0|income+age+edu+n1+n2+n3+race+home+lur|0,flabor)
```

上記のコマンドにより質的従属変数 h.choice に対して income,age,edu,…,lur という個人の属性を示す変数を説明変数としてロジット分析を行っている、結果は以下のように出力される。

```
> summary(result)
Call:
mlogit(formula = choice ~ 0 | income + age + edu + n1 + n2 +
    n3 + race + home + lur, data = flabor, method = "nr", print.level = 0)

Frequencies of alternatives:
```

```
      1       2       3       4       5       6
0.264636 0.119456 0.103489 0.141336 0.315789 0.055293

nr method
6 iterations, 0h:0m:3s
g'(-H)^-1g = 0.000349
successive fonction values within tolerance limits

Coefficients :
             Estimate  Std. Error  t-value  Pr(>|t|)
alt2        0.15452327  0.51548357   0.2998 0.7643574
alt3        0.19628869  0.54213647   0.3621 0.7173034
alt4       -0.84531999  0.50369453  -1.6782 0.0933004 .
alt5        1.83077912  0.41730633   4.3871 1.149e-05 ***
alt6       -0.75449779  0.69532506  -1.0851 0.2778770
alt2:income -0.00026096  0.00021319  -1.2241 0.2209209
alt3:income -0.00096086  0.00030384  -3.1624 0.0015649 **
alt4:income -0.00166020  0.00031509  -5.2690 1.372e-07 ***
alt5:income -0.00178525  0.00026085  -6.8441 7.696e-12 ***
alt6:income -0.00201437  0.00048994  -4.1115 3.932e-05 ***
alt2:age    -0.05220222  0.00731662  -7.1347 9.697e-13 ***
alt3:age    -0.05214643  0.00736760  -7.0778 1.465e-12 ***
alt4:age    -0.06633841  0.00691800  -9.5892 < 2.2e-16 ***
alt5:age    -0.07933691  0.00564324 -14.0588 < 2.2e-16 ***
alt6:age    -0.07869738  0.00942287  -8.3517 < 2.2e-16 ***

…

t5:home    0.63847315  0.12325079   5.1803 2.216e-07 ***
alt6:home    0.35831527  0.20349818   1.7608 0.0782759 .
alt2:lur    -0.04453642  0.02608895  -1.7071 0.0878037 .
alt3:lur    -0.11954816  0.02982513  -4.0083 6.116e-05 ***
alt4:lur    -0.09169399  0.02671970  -3.4317 0.0005998 ***
alt5:lur    -0.10937905  0.02186478  -5.0025 5.658e-07 ***
alt6:lur    -0.11590765  0.03908063  -2.9659 0.0030184 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -5133.5
McFadden R^2:  0.075028
Likelihood ratio test : chisq = 832.79 (p.value=< 2.22e-16)
```

Coefficients が係数の推定値である。ここで alt2,…,alt6 は定数項の推定値、alt2:income,…,alt6:income は income の係数の推定値、alt2:age,…,alt6:age は age の係数の推定値、等である。alt1 や alt1:income が無い事に注意しよう。これは識別性の問題によりすべての選択肢の効用の係数は識別できず、識別できるのは係数の差のみになるので、ここでは自動的に選択肢 1 の係数との差を推定している。推定値の横には標準誤差や t 値(推定値/標準誤差)も出力されている。解釈は回帰分析の時と同じである。

先ほどは自動的に選択肢 1 の係数との差をとったが、reflevel というオプションを使えば、どの選択肢の係数と差をとるかを指定する事もできる。例えば選択肢 2 の係数との差を推定したければ

```
> result = mlogit(h.choice~0|income+age+edu+n1+n2+n3+race+home+lur,flabor,reflevel="2")
```

とすればよい。結果は同様に summary でみられる。

```
> summary(result)

Call:
mlogit(formula = choice ~ 0 | income + age + edu + n1 + n2 +
    n3 + race + home + lur, data = flabor, reflevel = "2", method = "nr",
    print.level = 0)

Frequencies of alternatives:
      2       1       3       4       5       6
0.119456 0.264636 0.103489 0.141336 0.315789 0.055293

nr method
6 iterations, 0h:0m:2s
g'(-H)^-1g = 0.000349
successive fonction values within tolerance limits

Coefficients :
             Estimate  Std. Error  t-value  Pr(>|t|)
alt1       -1.5452e-01  5.1548e-01  -0.2998 0.7643574
alt3        4.1765e-02  6.1409e-01   0.0680 0.9457764
alt4       -9.9984e-01  5.7631e-01  -1.7349 0.0827567 .
alt5        1.6763e+00  5.0388e-01   3.3267 0.0008788 ***
 ...
alt4:lur    -4.7158e-02  3.0981e-02  -1.5222 0.1279704
alt5:lur    -6.4843e-02  2.6955e-02  -2.4056 0.0161455 *
alt6:lur    -7.1371e-02  4.2022e-02  -1.6984 0.0894303 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5133.5
McFadden R^2:  0.075028
Likelihood ratio test : chisq = 832.79 (p.value=< 2.22e-16)
```

4. mlogit による多項ロジットモデルの推定（説明変数が選択肢*j* にも依存している場合）

3 では説明変数が個人 i にのみ依存しているような場合の多項ロジットモデルの推定の仕方を説明した。一般には説明変数は選択肢 j にも依存している場合がある。すなわち、より一般的には選択肢 j からの効用 U_{ij} は

$$U_{ij} = \alpha_j + \beta X_{ij} + \gamma_j Z_i + \delta_j W_{ij} + \varepsilon_{ij}$$

のように表現する事ができる。この時、識別できるパラメーターは

$$\kappa_{jk} = \alpha_j - \alpha_k,\ \beta,\ \omega_{jk} = \gamma_j - \gamma_k,\ \delta_j$$

である(もう一つの資料参照)。いくつかのパラメーターについてはパラメーター間の差のみしか推定できないことに注意。このような場合にこれらのパラメーターを mlogit を用いてどのように推定するかを説明する。

4.1 データの準備

ここでは mlogit パッケージについている Fishing というデータを用いて説明する。まずデータを読み込

むために

```
> data("Fishing", package ="mlogit")
```

と入力する。すると mlogit パッケージについている Fishing というデータが読み込める。このデータを見ると

```
> head(Fishing, 5)
     mode price.beach price.pier price.boat price.charter catch.beach catch.pier catch.boat catch.charter income
1 charter     157.930     157.930    157.930       182.930      0.0678     0.0503     0.2601       0.5391 7083.332
2 charter      15.114      15.114     10.534        34.534      0.1049     0.0451     0.1574       0.4671 1250.000
3    boat     161.874     161.874     24.334        59.334      0.5333     0.4522     0.2413       1.0266 3750.000
4    pier      15.134      15.134     55.930        84.930      0.0678     0.0789     0.1643       0.5391 2083.333
5    boat     106.930     106.930     41.514        71.014      0.0678     0.0503     0.1082       0.3240 4583.332
```

となっている。これは Fishing mode の選択に関するデータであり mode が Fishing mode を表し、beach, pier, boat, charter の 4 つがある。また説明変数としてはまず個人 *i* にだけ依存している income と、個人 *i* とそれぞれの選択肢に依存している price (price.beach, price.pier, price.boat, price.charter) とcatch (catch.beach, catch.pier, catch.boat, catch.charter) がある。個人 *i* の選択肢 *j* (*j* = beach, pier, boat, charter) からの効用を

$$U_{ij} = \alpha_j + \beta X_{ij} + \gamma_j Z_i + \delta_j W_{ij} + \varepsilon_{ij}$$

とした時の関連するパラメーターを mlogit で推定してみよう。ここで X_{ij} を price、Z_i を income W_{ij} を catch に関する説明変数とする。X_{ij} の係数はそれぞれの j で共通である事に注意。

まず、3 の時と同様、データを mlogit で扱える形式に直す。

```
> Fish = mlogit.data(Fishing, shape ="wide", varying = 2:9, choice = "mode")
> head(Fish,9)
           mode    income     alt    price   catch chid
1.beach   FALSE  7083.332   beach  157.930  0.0678    1
1.boat     FALSE 7083.332    boat  157.930  0.2601    1
1.charter  TRUE   7083.332 charter 182.930  0.5391    1
1.pier     FALSE 7083.332    pier  157.930  0.0503    1
2.beach    FALSE 1250.000   beach  15.114   0.1049    2
2.boat     FALSE 1250.000    boat   10.534   0.1574    2
2.charter  TRUE   1250.000 charter  34.534   0.4671    2
2.pier     FALSE 1250.000    pier   15.114   0.0451    2
3.beach    FALSE 3750.000   beach  161.874  0.5333    3
…
```

ここで新たに varying というオプションが出てきているが、これは元のデータの何列目から何列目が選択肢に依存した説明変数かを示している。

4.2 mlogit による推定

このモデルを mlogit によって推定するには

```
> result.fish = mlogit(mode ~ price |income | catch, Fish)
```

と入力する。mlgoti の括弧の中は

質的従属変数(mode)~ X_{ij}とする変数(price)| Z_iとする変数(income)| W_{ij}とする変数(catch)，データの名前(Fish)

のように並べる。上の例では X_{ij}, Z_i, W_{ij} のような変数がそれぞれ一つずつあるが、例えば Z_i に属するような変数がない場合は

質的従属変数 ~ X_{ij}とする変数 | 0| W_{ij}とする変数，データの名前

X_{ij} に属するような変数がない場合は

質的従属変数 ~ 0 | Z_iとする変数 | W_{ij}とする変数，データの名前

のように 0 とする(3 節も参照)

Fish の分析結果は

```
> summary(result.fish)

Call:
mlogit(formula = mode ~ price | income | catch, data = Fish,
    method = "nr", print.level = 0)

Frequencies of alternatives:
  beach    boat charter    pier
0.11337 0.35364 0.38240 0.15059

nr method
7 iterations, 0h:0m:1s
g'(-H)^-1g = 2.54E-05
successive fonction values within tolerance limits

Coefficients :
                  Estimate  Std. Error  t-value  Pr(>|t|)
altboat          8.4184e-01  2.9996e-01   2.8065 0.0050080 **
altcharter       2.1549e+00  2.9746e-01   7.2443 4.348e-13 ***
altpier          1.0430e+00  2.9535e-01   3.5315 0.0004132 ***
price           -2.5281e-02  1.7551e-03 -14.4046 < 2.2e-16 ***
altboat:income    5.5428e-05  5.2130e-05   1.0633 0.2876612
altcharter:income -7.2337e-05  5.2557e-05  -1.3764 0.1687088
altpier:income   -1.3550e-04  5.1172e-05  -2.6480 0.0080977 **
altbeach:catch    3.1177e+00  7.1305e-01   4.3724 1.229e-05 ***
altboat:catch     2.5425e+00  5.2274e-01   4.8638 1.152e-06 ***
altcharter:catch  7.5949e-01  1.5420e-01   4.9254 8.417e-07 ***
altpier:catch     2.8512e+00  7.7464e-01   3.6807 0.0002326 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -1199.1
McFadden R^2:  0.19936
Likelihood ratio test : chisq = 597.16 (p.value=< 2.22e-16)
```

となる。income についての変数は beach に関する係数との差が推定されていることに注意。

練習問題

1．Fish データについて price と catch を共に W_{ij} のような変数として多項ロジットモデルを推定せよ。

2．Fish データについて price と catch を共に X_{ij} のような変数として多項ロジットモデルを推定せよ。