Rによる順序ロジットモデルの推定[†]

順序ロジット(ordered logit)モデルをRで推定する方法について説明する。

1．パッケージ ordinal のインスツール

順序ロジットモデルを推定するためにRのパッケージ ordinal 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。

「パッケージ」→「パッケージのインスツール…」→(適当なミラーサイトを選択:基本的にどこでもよい)→「ordinal」を選択→いろいろ表示されて、パッケージのインスツール完了。

次にコマンドウィンドウ(R Console)で

```
> library(ordinal)
```

と入力するとパッケージ ordinal 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(多項ロジットモデルの推定の時に出てきた) flabordata.txt を用いる(データについて詳しくは「R による多項ロジットモデルの推定」を参照)。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、h.choice が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 6 段階に分けたものである。数字が高いほど、より多くの就業時間を選択している事となり、1 は働いていないことを意味する(詳しくは「Rによる多項ロジットモデルの推定」を参照)。このデータに対して順序ロジットモデルを推定してみよう。そのためのRの関数は clm()である。

まず、上記のように読み込んだデータをさらに as.ordered というコマンドを用いて clm() で扱うデータの形式に直す。

[†]この資料は私のゼミおよび講義でRの使用法を説明するために作成した資料です。ホームページ上で公開しており、自由に参照して頂いて構いません。ただし、内容について、一応検証してありますが、間違いがあるかもしれません。間違いがあった場合でもそれによって生じるいかなる損害、不利益について責任は負いかねますのでご了承ください。

```
> h.choice.ord = as.ordered(flabordata$h.choice)
> head(h.choice.ord,10)
 [1] 5 2 5 1 6 1 1 3 5 1
Levels: 1 < 2 < 3 < 4 < 5 < 6
```

ここで flabordata$h.choice は flabordata というデータにある h.choice という変数であるという事を示している。

3. clm による順序ロジットモデルの推定

3.1 順序ロジットモデルの推定

前節で準備したデータに対して、順序ロジットモデルを推定する。以下のコマンドで推定する。

```
> result= clm(h.choice.ord~income+age+edu+n1+n2+n3+race+home+lur,
+ data=flabordata)
```

ここで「 …+lur, 」まで打ち込んだら Enter キーを押すと(R ではコマンドが長くなり 1 行で収まりそうにない場合は Enter キーを押すと次行に移動する)「 + 」が表示されて、そこから続きのコマンドを打ち込むことができるようになる。最後の data=flabordata は説明変数として入力された income や age などのデータがどこのデータかを指定している(ここで「data=」は必ずつける。これをつけないとうまくいかない)。(警告メッセージが出るが、ひとまずそれは無視する)上記のコマンドにより(順序付きの)質的従属変数 h.choice.ord 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順序ロジットモデルの推定を行っている、結果は以下のように出力することができる。

```
> summary(result)
formula:
h.choice.ord ~ income + age + edu + n1 + n2 + n3 + race + home + lur
data:    flabordata

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  3382 -5224.34 10476.69 6(0)  9.45e-08 4.7e+07

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
income -0.001145   0.000158  -7.247 4.27e-13 ***
age    -0.050316   0.003562 -14.128  < 2e-16 ***
edu     0.160564   0.014696  10.926  < 2e-16 ***
n1     -0.843816   0.048546 -17.382  < 2e-16 ***
n2     -0.259377   0.037704  -6.879 6.02e-12 ***
n3      0.102520   0.063334   1.619    0.106
race    0.314242   0.073203   4.293 1.76e-05 ***
home    0.435912   0.076323   5.711 1.12e-08 ***
lur    -0.073537   0.014150  -5.197 2.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
```

```
1|2 -1.89713    0.26160  -7.252
2|3 -1.25438    0.26058  -4.814
3|4 -0.75680    0.26021  -2.908
4|5 -0.09462    0.25984  -0.364
5|6  2.38765    0.26654   8.958
```

Coefficients が係数の推定値である。Threshold coefficients は選択肢を分ける境界値の推定値である。例えば 1|2 の右横の値が選択肢 1 と 2 を分ける境界の推定値となる。推定値の横には標準誤差、z 値(推定値/標準誤差)、z 値の P 値、も出力されている。解釈は回帰分析の時と同じである。係数の推定値を見てみると符号については全て直観とあっている。例えば income の係数はマイナスだがこれは夫その他の収入が多ければより少ない就業時間を選ぶという事で直観的にあっている。

また、いわゆるあてはめ値(fitted value)は

```
> fitted(result)
```

によって得られる。これは実際に選択された選択肢が取られる確率をモデルから計算したものである。

先ほどはまずデータを clm で扱えるように変換し、変換したデータに対して分析をしたが、これらは以下のように入力すれば一度にできる

```
> result = clm(as.ordered(h.choice)~income+age+edu+n1+n2+n3+race+home+
+ lur,data=flabordata)
```

出力結果は同じである。

3.2 警告メッセージについて

先ほどの推定において

```
警告メッセージ:
(2) Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
```

という警告メッセージが出たが、これは説明変数間の変動の差が極端に大きかったり(例えばある説明変数の分散が 10000 であるのに対して、他のある説明変数の分散が 1 であるような場合)すると起こるようである。上記のデータでは income の変動が他の変数の変動に比べて非常に大きいので、income のデータを rescale して 100 分の 1 にしたものを使用してみよう(0.01 を掛ける)

新たに

```
> income2=0.01*flabordata$income
```

というデータを作る。これを用いて

```
> result = clm(h.choice.ord~income2+age+edu+n1+n2+n3+race+home+
+ lur,data=flabordata)
```

とする。今回は警告メッセージが出ない。結果は

```
> summary(result)
formula: as.ordered(h.choice) ~ income2 + age + edu + n1 + n2 + n3
+ race + home + lur
data:    flabordata

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  3382 -5224.34 10476.69 6(0)  1.23e-08 5.6e+05

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
income2 -0.114524   0.015804  -7.247 4.27e-13 ***
age     -0.050316   0.003562 -14.128  < 2e-16 ***
edu      0.160564   0.014696  10.926  < 2e-16 ***
n1      -0.843816   0.048546 -17.382  < 2e-16 ***
n2      -0.259377   0.037704  -6.879 6.02e-12 ***
n3       0.102520   0.063334   1.619    0.106
race     0.314242   0.073203   4.293 1.76e-05 ***
home     0.435912   0.076323   5.711 1.12e-08 ***
lur     -0.073537   0.014150  -5.197 2.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2 -1.89713    0.26160  -7.252
2|3 -1.25438    0.26058  -4.814
3|4 -0.75680    0.26021  -2.908
4|5 -0.09462    0.25984  -0.364
5|6  2.38765    0.26654   8.958
```

となる。ほとんど変わらない(警告メッセージが出てもたいていの場合はうまく推定できているのでそこまで気にする必要もない?)。income2 の係数の推定値と標準誤差が(もともとの変数 income の係数と標準誤差の)100 倍になることに注意。

3.3 順序プロビットモデルの推定

clm 関数を用いて順序プロビットモデルを推定するには引数として「 link = "probit" 」を追加する。例えば先ほどの場合は

```
> result = clm(h.choice.ord~income2+age+edu+n1+n2+n3+race+home+
+ lur,data=flabordata,link="probit")
```

とする。

プロビットモデルとロジットモデルの推定値の間には**おおよそ**

$$\hat{\beta}^{\text{probit}} \approx \frac{\hat{\beta}^{\text{logit}}}{\pi/\sqrt{3}}$$

という関係がある（あくまでもおおよその関係で正確には成り立たない）。ここで $\hat{\beta}^{\text{probit}}$ はプロビットモデルにおける係数の推定値、$\hat{\beta}^{\text{logit}}$ はロジットモデルによる係数の推定値である。

練習問題

1．flabordata について上記のモデルを推定し、上記のプロビットモデルとロジットモデルのおおよその関係がどの程度成り立っているかを確認せよ（ヒント：プロビットモデルの推定結果を presult、ロジットモデルの推定結果を lresult とし、

```
> (pi/sqrt(3))*presult$coefficients
```

が lresult$coefficients とどの程度同じかを確認する）。