R による二項ロジットモデルの推定[†]

1. 分析の準備

毎回の事になるが、まず作業ディレクトリをデータが置いてあるディレクトリに変更する。
(「ファイル」→「ディレクトリの変更」で変更)。作業ディレクトリを移動したら、次にそこにデータがあるかどうかを

```
> dir()
```

で確認する。そこにおいてあるファイル一覧が出力される。今回使用するデータは dpdata.txt にある死刑制度の有無に関するアメリカの州のデータである。

```
> dpdata = read.table("dpdata.txt", header=T, skip=8)
```

で dpdata.txt のデータを読み込み dpdata という名前をつける。

```
> head(dpdata,3)
   D1     M    PC   T    Y    NW D2
1   1 19.25 0.204  47 1.10 0.321  1
2   1  7.53  0.327  58 0.92 0.224  1
3   1  5.66  0.401  82 1.72 0.127  0
```

というデータが入っている(各変数についての説明は dpdata.txt 参照)

2. ロジットモデルの推定

D1 を質的従属変数、それ以外を説明変数とするロジットモデルを推定する。glm()関数を用いる。

```
> results=glm(D1~M+PC+T+Y+NW+D2,family=binomial(link="logit"),data=dpdata)
```

と入力する。結果は results という変数にまとめられる。うまくいけば

```
> results=glm(D1~M+PC+T+Y+NW+D2,family=binomial(link="logit"),data=dpdata)
 警告メッセージ:
 glm.fit: 数値的に 0 か 1 である確率が生じました
>
```

と表示される。また dpdata の中で D1 以外の全ての変数を説明変数として使っているが、この場合、

```
> results=glm(D1~.,family=binomial(link="logit"),data=dpdata)
```

と説明変数の入力を “.” を入力する事によって省略する事ができる。推定結果は

```
> summary(results)
```

[†]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。

によって見る事ができる。結果は

```
Call:
glm(formula = D1 ~ ., family = binomial(link = "logit"), data = dpdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4832   0.0000   0.0000   0.0384   2.1737

Coefficients:
            Estimate Std. Error   z value Pr(>|z|)
(Intercept) -22.65966   13.81364  -1.640   0.1009
M             0.85730    1.00846   0.850   0.3953
PC           -9.86641    6.97014 -1.416   0.1569
T             0.03165    0.01582   2.000   0.0455 *
Y             8.39533    5.81075   1.445   0.1485
NW          126.17760   70.09161  1.800   0.0718 .
D2           18.87656 3399.69662   0.006   0.9956
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.584  on 43  degrees of freedom
Residual deviance: 14.160  on 37  degrees of freedom
AIC: 28.160

Number of Fisher Scoring iterations: 20
```

のように出力される。Estimate がそれぞれの説明変数の係数の推定値、Std. Error が推定値の標準誤差、z value というのは回帰分析の t value に対応するもので帰無仮説（ここでは係数が 0 という帰無仮説）のもとで標準正規分布に従う変数、Pr(> |z|)は z の P 値（帰無仮説のもとで z の絶対値が z value の値を超える確率）である。AIC は赤池情報量基準（Akaike Information Criteria）と呼ばれるものでモデルの当てはまりの良さをみる尺度。

係数の推定値を用いて D1 が 1 になる確率を計算（予測）したものは

```
> results$fitted
```

と入力する事によって求まる。その他の結果の見方については help(glm)で確認できる。

3. 限界確率効果の計算

限界確率効果をRで計算する。glm()関数はこれを自動的に計算して出力してくれないので、自分で計算式から計算しないといけない。ロジットモデルの場合は比較的簡単に計算できるので、以下ではロジットモデルの場合のみ説明する。

$\Lambda(x)$ をロジスティック分布の分布関数とすると、ロジットモデルの限界確率効果は

$$MPE_{ij} = \Lambda(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})[1 - \Lambda(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})]\,\beta_j$$

と計算できる。ここで $\Lambda(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$ はロジットモデルの場合、1 になる確率に他ならない。
よって、例えば $i = 4$, $j = 3$ に対して MPE_{ij} は

```
> A = results$fitted
> coef = coef(results)  (coef(results)は係数のベクトルを返す)
> MPE = A*(1-A)*coef[3] ( "*" は行列(ベクトル)の要素ごとの積を表す)
> MPE43 = MPE[4] ( MPE という行列の 4 番目の要素を取り出す)
```

などのように計算できる。答えは

```
> MPE43
      4
-0.01682834
```

となる。平均限界効果は MPE の平均をとった

```
> AMPE = mean(MPE)
```

によって計算できる。

4. プロビットモデルの推定

プロビットモデルもまったく同様に推定できる。

```
> Presults = glm(D1~.,family=binomial(link= "probit"), data=dpdata)
```

と結果に Presults という名前を付けよう。結果は

```
>summary(Presults)

Call:
glm(formula = D1 ~ ., family = binomial(link = "probit"), data = dpdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.49036  0.00000  0.00000  0.00546  2.09611

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.503670   7.596508  -1.778   0.0755 .
M             0.474984   0.556296   0.854   0.3932
PC           -5.504881   3.792651  -1.451   0.1467
T             0.017945   0.008169   2.197   0.0280 *
Y             5.115795   3.245519   1.576   0.1150
NW           72.322867  38.571836   1.875   0.0608 .
D2            6.264910 660.258684   0.009   0.9924
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.584  on 43  degrees of freedom
Residual deviance: 14.109  on 37  degrees of freedom
AIC: 28.109

Number of Fisher Scoring iterations: 20
```

のようになる。link = "logit" が link = "probit" に変わるだけである。

練習

同じことを schooldata.txt について SCHOOL を被説明変数 EDU, INC を説明変数としてやってみる。