統計解析論特殊講義

Gretl を使った計量経済学６

中川　満

今回は，離散選択モデルを学習する．

教科書 p.196 の例にあるように，家を所有する確率が所得，勤続年数，年齢で決まるようなモデルを考える．数式で表すと，

$$\text{確率} = f(\alpha + \beta_1 \text{所得} + \beta_2 \text{勤続年数} + \beta_3 \text{年齢})$$

ここで，$f(t)$は t の増加関数である．問題は，$f(t)$の関数形である．計量経済学では，以下の二つの関数形を想定している．$t = \alpha + \beta_1 \text{所得} + \beta_2 \text{勤続年数} + \beta_3 \text{年齢}$ とすると，① probit モデルの場合，

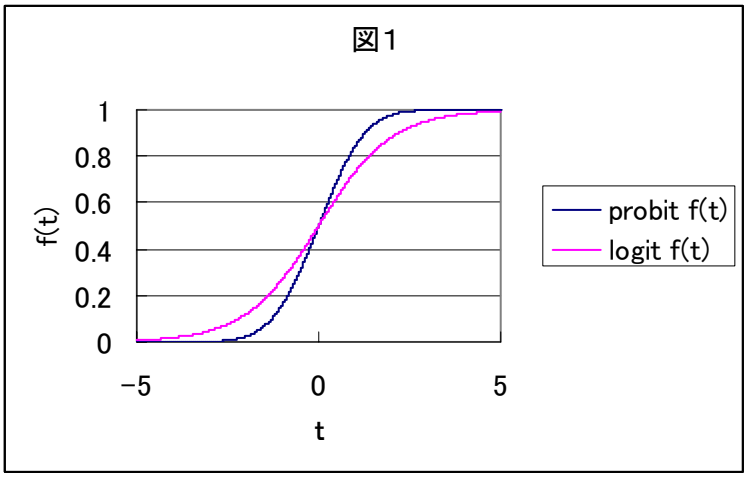
$$f(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt$$

であり，②logit モデルの場合，

$$f(t) = \frac{\exp(t)}{1+\exp(t)}$$　（ロジスティック関数と呼ばれる）

である．両方の関数のグラフを書くと図１のようになる．

図1



どちらのモデルも$\alpha + \beta_1 \text{所得} + \beta_2 \text{勤続年数} + \beta_3 \text{年齢} = 0$となるところで，持ち家を所有する確率が 0.5 になると考えるモデルである．そして，（$\alpha + \beta_1 \text{所得} + \beta_2 \text{勤続年数} + \beta_3 \text{年齢}$）が大きくなれば，持ち家所有の確率が高まるのである．もし，β_k がプラスであれば，その変数の上昇が持ち家確率の上昇を促し，マイナスであれば，その変数の上昇が持ち家確率の下落を促す．さらに，その変数が１単位大きくなることによる持ち家所有確率の上昇分も計算できる．ただし，これは非線形関数なので，値が一定ではない．従って，各変数の標本平均値を想定して，この値を計算する．これを，SLOPE とよぶ．

定数項は，$\alpha = -\beta_1$所得 $-\beta_2$勤続年数 $-\beta_3$年齢 において持ち家を所有する確率が 0.5 になるということから解釈できる．

１．logit モデル

まず，logit モデルを推定してみよう．

① gretl を起動する．

② メニューバーの「ファイル」→「データを開く」→「Sample file」を左クリックする．

③ 出てきた「gretl: data flies」ウィンドウの「Greene」タブを左クリックする．

④ リストの中の「greene19_1」をみつける．

⑤ このデータの説明を見てみよう．この行を右クリックし，ポップアップしたリストの中から「Info」を左クリックする．
以下のような記述がでてくる．

```
Data used to study program effectiveness.
See William Greene, Econometric Analysis (4 ed.), Example 19.1.

GPA = grade point average
TUCE = test score on economics test
PSI = participation in program
GRADE = grade increase (1) or decrease (0) indicator

This file included with gretl by kind permission of William Greene.
```

読むと，経済学の新しい教育プログラムの導入の効果を検証するためのデータであることが分かる．

GPA：GPA ポイント
（例えば A(特優)=4, B(優)=3, C(良)=2, D(可)=1 をそれぞれの単位と掛け合わせ合計し，それを総取得単位数で割ったもの．）

TUCE:教育プログラムの事前試験成績（経済学）

PSI:教育プログラムに参加したかどうか（１なら参加，０なら不参加）のダミー変数

GRADE:事前試験の成績よりアップしたかどうかを示すダミー（１ならアップ，０なら等しいかダウン）

原論文を読んだのではないが，おそらく PSI=0 の人たちは旧式のプログラムを受けたのだろう．

⑥　「gretl: data flies」ウィンドウにもどり，リストの中の「greene19_1」をダブルクリックする．

⑦　まず比較対象として，OLS（最小二乗法）を行う．この場合，理論的には，Heteroskedasticity があると考えられるので，Heteroskedasticity-Robust な標準誤差の補正を行うことにする. gretl メインウィンドウのメニューバーの「モデル」→「Ordinary Least Squares」を左クリックする．

⑧　でてきた「gretl: Specify model」ウィンドウの「dependent variable」の入力ボックスに「GRADE」を「Choose ->」ボタンを左クリックして指定する．

⑨　「independent variable」には残りの変数を指定する．

⑩　このウィンドウの左下の「Robust standard errors」のチェックボックスを左クリックして，チェックを入れる．

⑪　「OK」ボタンを左クリックする．

⑫　出力の解釈

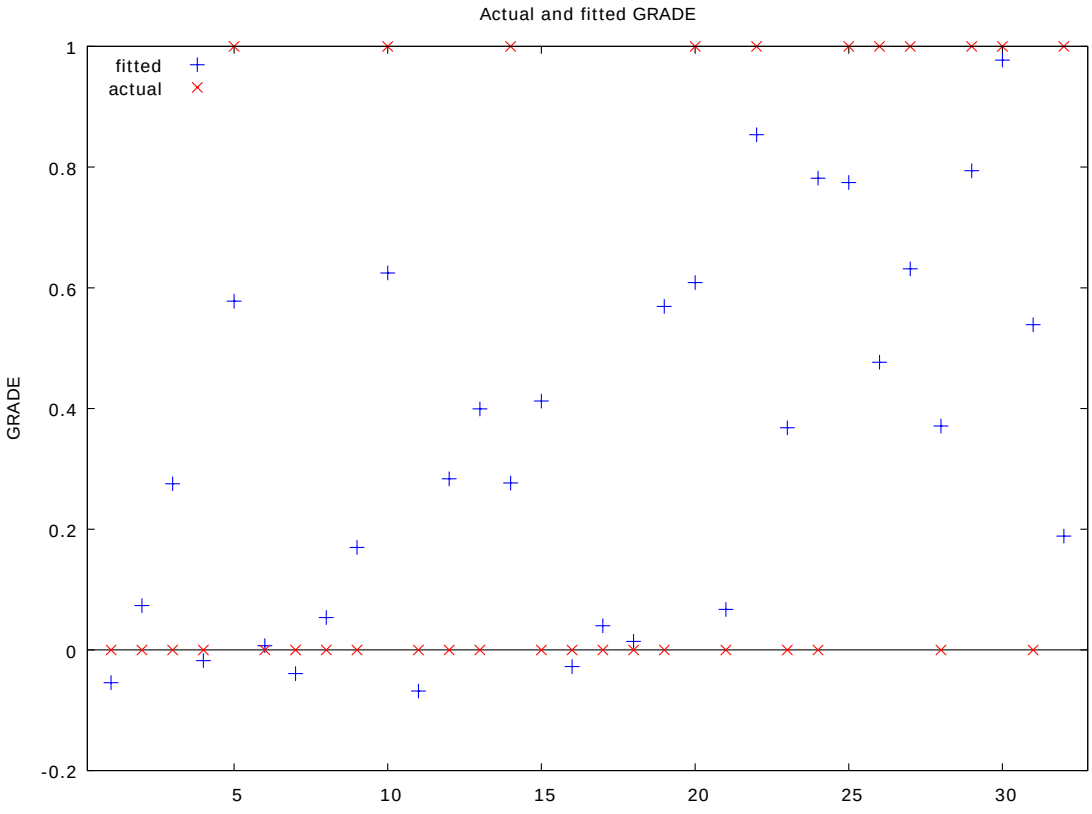
```
Model □: OLS estimates using the 32 observations 1-32

Dependent variable: GRADE

Heteroskedasticity-robust standard errors, variant HC1

     VARIABLE      COEFFICIENT       STDERROR     T STAT   P-VALUE

  const              -1.49802         0.497553    -3.011   0.00547 ***
  GPA                 0.463852        0.151097     3.070   0.00472 ***
  TUCE                0.0104951       0.0174117    0.603   0.55152
  PSI                 0.378555        0.150361     2.518   0.01781 **

  Mean of dependent variable = 0.34375
  Standard deviation of dep. var. = 0.482559
  Sum of squared residuals = 4.21647
  Standard error of residuals = 0.388057
  Unadjusted R-squared = 0.4159
  Adjusted R-squared = 0.353318
  F-statistic (3, 28) = 11.0999 (p-value = 5.65e-005)
  Log-likelihood = -12.9782
  Akaike information criterion (AIC) = 33.9565
  Schwarz Bayesian criterion (BIC) = 39.8194
  Hannan-Quinn criterion (HQC) = 35.8999
```

```
Excluding the constant, p-value was highest for variable 2 (TUCE)
```

p値をみると，有意に0と異なる係数は，GPA と PSI であり，係数はプラスである．つまり，プログラムを受ける前の経済学の知識と関係なしに，一般的な成績水準と新しいプログラムを受けたという２点が前の試験の得点を超える確率を上昇させることになる．例えば，GPA が１ポイント上がると，（つまり，すべての試験で１ランク上の成績をとると，），前の試験の得点を超える確率を 0.45 上昇させ，このプログラムを受けたことは，前の試験の得点を超える確率を 0.38 上昇させることになる．

⑬　次に，OLS のモデルのウィンドウ「gretl:model X（X はそれぞれの人で異なる）」のメニューバーの「Graphs」→「Fitted, actual Plot」にカーソルを合わせて，「By observation number」を左クリックする．十字の点は，それぞれの観測値，すなわち，それぞれの学生に関するこのモデルの予測確率を示す. 例えば, 予測確率が 0.8 近いのに，実際は成績が上がっていない学生が，24 番の学生である．逆に，予測確率が低いのに火星席が上がっている場合もある．さらによくこの結果のグラフをみると，この線型モデルの問題点が明らかになる．予測確率が負のデータが 1,4 などの 4 人ほどある．確率は非負であるから，このモデルの無理が明らかになる．

図２



⑭ gretl メインウィンドウのメニューバーの「モデル」→「Nonlinear Models」→「Logit」を左クリックする．

⑮ 「gretl: specify model」ウィンドウが開く．「dependent variable」の入力ボックスに「GRADE」を，「independent variable」には残りの変数を指定する．さらに，のウィンドウの左下の「Robust standard errors」のチェックボックスを左クリックして，チェックを入れる．そして，「OK」ボタンを左クリックする．

⑮ 結果の評価

以下のような出力結果が得られる．

```
Convergence achieved after 6 iterations

Model 5: Logit estimates using the 32 observations 1-32
Dependent variable: GRADE
QML standard errors

     VARIABLE      COEFFICIENT       STDERROR      T STAT      SLOPE
                                                             (at mean)
  const             -13.0213          5.19759     -2.505
  GPA                2.82611          1.26755      2.230      0.533859
  TUCE               0.0951577        0.117922     0.807      0.0179755
  PSI                2.37869          0.964419     2.466      0.449339

  Mean of GRADE = 0.344
  Number of cases 'correctly predicted' = 26 (81.3%)
  f(beta'x) at mean of independent vars = 0.189
  McFadden's pseudo-R-squared = 0.374038
  Log-likelihood = -12.8896
  Likelihood ratio test: Chi-square(3) = 15.4042 (p-value 0.001502)
  Akaike information criterion (AIC) = 33.7793
  Schwarz Bayesian criterion (BIC) = 39.6422
  Hannan-Quinn criterion (HQC) = 35.7227

          Predicted
             0    1
  Actual 0  18    3
```

```
        1   3    8
```

```
const              -13.0213          5.19759      -2.505
```

定数項の値は−13.0 で，推定値の標準誤差は 5.20，t値は−2.51 なので，有意である．この場合，誤差項が正規分布ではないので，t値の絶対値と 1.96 を比較し，絶対値が大きければ係数値は0と有意に異なる．そうでなければ，0と有意に異ならない．

```
GPA                 2.82611          1.26755       2.230      0.533859
```

GPA の係数は，2.83 で，標準誤差は 1.27，t値は 2.23 で，説明変数の平均における GPA の 1 点の増加による経済学試験の点数を改善させる確率の増加は 0.53 である．では，この平均値はどんな値なのか，それは，前の方のプリントの「Summary statistics」の出し方を参考にしてほしい．以下が当該の平均値である．

```
GPA            3.1172
TUCE          21.938
PSI            0.43750
```

そして，この値を OLS での推定量である 0.463852 と比較すると，logit モデルの方が OLS よりやや大きく推定されていることが分かる．logit の方がより正確だと考えられるので，GPA に表された学力の高さの貢献は OLS が示すものより大きいと言うことになる．

```
TUCE                0.0951577        0.117922       0.807      0.0179755
```

これは，t値の絶対値が 1.96 より小さいので，有意ではない．つまり，教育プログラムにかかる以前に持っていた経済学に関する知識は，得点を改善させることに貢献しなかったと言うことである．

```
PSI                 2.37869          0.964419       2.466      0.449339
```

この係数も有意で，PSI プログラムに参加することで，点数が改善する確率を 0.44 上げることができることを示している．OLS の推定値と比較すると，0.378555 から 0.45 となって，OLS の改善効果の推定値は過小であることが分かる．

```
Number of cases 'correctly predicted' = 26 (81.3%)
```

これは，26 個の観測値についてモデルの予測と実際の予測の結果が一致したことを示す．この場合，モデルの予測した確率が 0.5 以上をテストの点数が上がると予測したとし，確率 0.5 以下を上がらないと予測したと見なして，的中を判断するのである．かなりいいモデルだといえる．

```
f(beta'x) at mean of independent vars = 0.189
```

これは，先ほどの SLOPE の値の評価に用いる．つまり，各説明変数の標本平均で評価したときの点数改善確率で，これに先ほどの SLOPE の値を足すと説明変数の値が平均から1上がったときの，限界効果をみることができる．

```
 McFadden's pseudo-R-squared = 0.374038
```

このモデルの説明力の指標で，1に近づけば近づくほど，説明力が高い．

```
Likelihood ratio test: Chi-square(3) = 15.4042 (p-value 0.001502)
```

これは，全ての説明変数（定数項は除く）の係数のうちのいずれかが０かどうかを判定する．ｐ値が 0.05 より小さければ，説明変数の係数のうちいずれかはは有意（５％水準で）に０と異なる．

```
         Predicted
            0    1
  Actual 0  18    3
         1   3    8
```

これは，予測と実際を 2×2 の分割表で表したものである．これは，予測に偏りがあるかどうかをみるのに役立つ．たとえば，点数の上昇を予測しているときは，的中確率が高いが，下落の予測の時は，的中確率が低いなどの情報がわかる．この場合は，点数の下落が予測されるときは，外れの確率が１／７，上昇が予測されるときには，外れの確率は３／１１であることがわかる．

２．probit モデル

① gretl メインウィンドウのメニューバーの「モデル」→「Nonlinear Models」→「Probit」を左クリックする．「gretl: specify model」ウィンドウが開く．「dependent variable」の入力ボックスに「GRADE」を，「independent variable」には残りの変数を指定する．さらに，のウィンドウの左下の「Robust standard errors」のチェックボックスを左クリックして，チェックを入れる．そして，「OK」ボタンを左クリックする．

② 出力の評価

出力形式は logit と同じであるので，判断は，logit のところと同じである．

```
Convergence achieved after 7 iterations


Model 1: Probit estimates using the 32 observations 1-32
Dependent variable: GRADE
QML standard errors

     VARIABLE      COEFFICIENT       STDERROR      T STAT      SLOPE
                                                             (at mean)
  const              -7.45232         2.54427      -2.929
  GPA                 1.62581        0.651510      2.495      0.533347
  TUCE               0.0517288       0.0691327     0.748      0.0169697
  PSI                 1.42633        0.532765      2.677      0.467908
```

```
Mean of GRADE = 0.344
Number of cases 'correctly predicted' = 26 (81.3%)
f(beta'x) at mean of independent vars = 0.328
McFadden's pseudo-R-squared = 0.377478
Log-likelihood = -12.8188
Likelihood ratio test: Chi-square(3) = 15.5459 (p-value 0.001405)
Akaike information criterion (AIC) = 33.6376
Schwarz Bayesian criterion (BIC) = 39.5006
Hannan-Quinn criterion (HQC) = 35.581

        Predicted
          0    1
Actual 0  18    3
       1   3    8
```

③　probit か logit か？

変数選択の場合のように AIC,BIC の最小化でモデル選択を行う．

OLS

```
Akaike information criterion (AIC) = 33.9565
Schwarz Bayesian criterion (BIC) = 39.8194
Hannan-Quinn criterion (HQC) = 35.8999
```

logit

```
Akaike information criterion (AIC) = 33.7793
Schwarz Bayesian criterion (BIC) = 39.6422
Hannan-Quinn criterion (HQC) = 35.7227
```

probit

```
Akaike information criterion (AIC) = 33.6376
Schwarz Bayesian criterion (BIC) = 39.5006
Hannan-Quinn criterion (HQC) = 35.581
```

上記のように，いずれの情報基準も probit を支持している．

３．ダミー変数について

・すべての観測値についてダミー変数の合計＝１となる場合，定数項を含む回帰を行ってもうまくいかない．ダミー変数のうち一つを取り除くこと．

・定数項を含まない回帰をやるには，「gretl: specify model」ウィンドウの「independent variable」の中の const を左クリックした後，「<-Remove」ボタンを左クリックする．これ

で，説明変数から定数項が外れた．その後，「OK」ボタンを左クリックする．

４．課題について

締め切り：８月８日１７：００まで

提出先：経済学部事務室

提出形態：レポートを紙に印刷，または，レポートファイル（MS-Word 形式か PDF 形式）を CD-R に格納

課題は７月２３日に中川の HP に掲示.

(URL:http://ramsey.econ.osaka-cu.ac.jp/~Nakagawa/07sasl.htm)

注意

掲示の手順に従って，まず，データをダウンロードし，その後，gretl 上でそのデータが各人毎異なるデータになるような操作をしてもらいます．その後，分析してください．その手順は HP 上に掲示します．注意深く読んで実行してください.

７月２６日までに，全てのプリントについて最終版をアップします．それを参考にしながら課題をやってください.