MLE

Ryuya Ko

12/24/2019

目的

- 質的データの分析に際して、最尤推定を導入する
- 最尤推定の大まかな概念を理解する
- 最尤推定の実装方法を理解する

最尤推定 (Maximum likelihood estimation)

カテゴリカルな値のみを取る質的データを扱う際、線形回帰モデルをそのまま適用するのは不適切なケースが多い。

代表的な例として、`Yes/1` または `No/0` の 2 値のみを取るような被説明変数を持つモデルが考えられる。

- 政府の施策への参加や、出産など
- 線形確率モデルを適用しても、現実とマッチしない場合がある

このような 2 値を取るデータを分析するモデルとして、トービットモデルやプロビットモデルを主に用いる。その際の推定手法として、最尤推定 (Maximum likelihood estimation, MLE) を用いることが多い。

尤度関数

求めるパラメータを θ, i 番目の被説明変数を $y_i(= 0, 1)$ とおく。 n 個の観測値が $i.i.d$ のとき、その joint density は次のように書ける:

$$f(y_1, \dots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \equiv L(\theta|y).$$

$L(\theta|y)$ を尤度関数と呼ぶ。 θ に条件づけたときの joint density を θ についての関数と見なしていることに注意したい。

直観的には、観測値の下で尤度関数を最大化するようなパラメータの推定値は、真のパラメータ (θ_0) にもっとも近いように思われる。実際、いくつかの正則条件 (regularity condition) のもとで、MLE 推定量 $\widehat{\theta}$ は次の性質を持つ:

- 一致性: $\hat{\theta} \xrightarrow{p} \theta_0$ as $n \to \infty$
- 漸近正規性: $\sqrt{n}(\hat{\theta} - \theta_0) \overset{d}{\sim} N(0, (I(\theta_0))^{-1})$. ここで、

$$I(\theta_0) = -E_0[\partial^2 \ln L / \partial\theta_0 \partial\theta']$$

- 漸近効率性: $\hat{\theta}$ は漸近的に効率的であり, クラメール=ラオの下界を達成する.

詳細な証明などは Greene の “Econometric Analysis” の 14 章を参照のこと。

プロビット・トービットモデル

(板書)

R を用いた実装

~~R の組み込みパッケージ stats4 に mle 関数が格納されている。~~ 今回はひとまず使わない方針で進める。

```
setwd('/Users/LOng/econ/shimotsu_seminar/subsemi/introductory_econometrics')
library(dplyr);library(wooldridge)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data("catholic");help("catholic")
head(catholic)
```

```
##       id read12 math12 female asian hispan black motheduc fatheduc  lfaminc
## 1 124902  61.41  49.77      0     0      0     0       14       12 10.30895
## 2 124915  58.34  59.84      0     0      0     0       14       14 10.30895
## 3 124916  59.33  50.38      1     0      0     0       14       11 10.30895
## 4 124932  49.59  45.03      1     0      0     0       12       14 10.30895
## 5 124944  57.62  54.26      1     0      0     0       12       12 10.65726
## 6 124947  52.53  56.73      1     0      0     0       12       11 11.04292
```

```
##   hsgrad cathhs parcath
## 1      1      0       1
## 2      1      0       1
## 3      1      0       1
## 4      1      0       1
## 5      1      0       1
## 6      1      0       1
```

```
str(catholic)
```

```
## 'data.frame':    7430 obs. of  13 variables:
##  $ id      : int  124902 124915 124916 124932 124944 124947 124966 124968 124972 1249
##  $ read12  : num  61.4 58.3 59.3 49.6 57.6 ...
##  $ math12  : num  49.8 59.8 50.4 45 54.3 ...
##  $ female  : int  0 0 1 1 1 1 1 0 0 0 ...
##  $ asian   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hispan  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ black   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ motheduc: num  14 14 14 12 12 12 14 14 14 14 ...
##  $ fatheduc: num  12 14 11 14 12 11 14 14 14 12 ...
##  $ lfaminc : num  10.3 10.3 10.3 10.3 10.7 ...
##  $ hsgrad  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cathhs  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ parcath : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(catholic)
```

```
##        id              read12          math12          female      
##  Min.   : 124902   Min.   :29.15   Min.   :29.50   Min.   :0.0000  
##  1st Qu.:2424049   1st Qu.:44.52   1st Qu.:45.02   1st Qu.:0.0000  
##  Median :4592442   Median :53.08   Median :52.53   Median :1.0000  
##  Mean   :4589838   Mean   :51.77   Mean   :52.13   Mean   :0.5174  
##  3rd Qu.:7241106   3rd Qu.:59.47   3rd Qu.:59.86   3rd Qu.:1.0000  
##  Max.   :7979086   Max.   :68.09   Max.   :71.37   Max.   :1.0000  
##
##      asian             hispan           black            motheduc    
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   : 8.00  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:12.00  
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :14.00  
##  Mean   :0.05168   Mean   :0.1035   Mean   :0.07066   Mean   :13.36  
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:14.00
```

```
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :18.00
##
##     fatheduc        lfaminc           hsgrad           cathhs
##  Min.   : 8.00   Min.   : 6.215   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:12.00   1st Qu.:10.021   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :14.00   Median :10.309   Median :1.0000   Median :0.00000
##  Mean   :13.67   Mean   :10.353   Mean   :0.9303   Mean   :0.06083
##  3rd Qu.:16.00   3rd Qu.:10.657   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :18.00   Max.   :12.346   Max.   :1.0000   Max.   :1.00000
##                                   NA's   :1460
##     parcath
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3459
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

```
prmodel <- glm(cathhs~read12+math12+female+asian+hispan+black,
             family=binomial(link="probit"), data=catholic)
summary(prmodel)
```

```
##
## Call:
## glm(formula = cathhs ~ read12 + math12 + female + asian + hispan +
##     black, family = binomial(link = "probit"), data = catholic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5992  -0.3960  -0.3370  -0.2797   2.8203
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.730877   0.162363 -16.820  < 2e-16 ***
## read12       0.006706   0.003709   1.808   0.0705 .
## math12       0.016140   0.003712   4.348 1.37e-05 ***
## female      -0.112225   0.047956  -2.340   0.0193 *
## asian       -0.129072   0.109728  -1.176   0.2395
## hispan       0.154850   0.076101   2.035   0.0419 *
```

```
## black        0.155437   0.094608   1.643   0.1004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3406.8  on 7429  degrees of freedom
## Residual deviance: 3333.0  on 7423  degrees of freedom
## AIC: 3347
##
## Number of Fisher Scoring iterations: 5
```

```
lgmodel <- glm(cathhs~read12+math12+female+asian+hispan+black,
               family=binomial(link="logit"), data=catholic)
summary(lgmodel)
```

```
##
## Call:
## glm(formula = cathhs ~ read12 + math12 + female + asian + hispan +
##     black, family = binomial(link = "logit"), data = catholic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6191  -0.3939  -0.3355  -0.2810   2.7907
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.219877   0.350682 -14.885  < 2e-16 ***
## read12       0.013710   0.007877   1.740   0.0818 .
## math12       0.033981   0.007868   4.319 1.57e-05 ***
## female      -0.230332   0.100454  -2.293   0.0219 *
## asian       -0.280691   0.233411  -1.203   0.2291
## hispan       0.326161   0.156961   2.078   0.0377 *
## black        0.323166   0.198829   1.625   0.1041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3406.8  on 7429  degrees of freedom
```

```
## Residual deviance: 3333.8  on 7423  degrees of freedom
## AIC: 3347.8
##
## Number of Fisher Scoring iterations: 6
```

```
catholic[, 'const'] <- 1
y <- catholic[, 'cathhs']
X <- catholic[, c('const', 'read12', 'math12',
                  'female', 'asian', 'hispan', 'black')]
```

```
# function(beta, obs){
#   beta %*%
# }
```