ДЗ-05 **Сдать 20.04.2023 до 23:50**

на почту islabolitskiy@hse.ru с темой письма «метрика_фамилия_дз…».

Например, «метрика_петров_дз05

Задача 1

Пусть есть логит-модель $P(y_i=1)=\Lambda(x_i'\beta)$

(а) Выпишите условия первого порядка для метода максимального правдоподобия.

(б) Поскольку y_i принимает 2 значение: 0 и 1, то $E(y_i)=1\cdot P(y_i=1)+0\cdot P(y_i=1)=P(y_i=1)=\Lambda(x_i'\beta)$, то $y_i=\Lambda(x_i'\beta)+\varepsilon_i$, где $E\varepsilon_i=0$. Применим нелинейный метод наименьших квадратов:

$g(\beta)=\sum_{i=1}^{n}(y_i-\Lambda(x_i'\beta))^2\xrightarrow[\beta]{}\min$. Выпишите условия первого порядка.

(в) Совпадают ли уравнения (а) и (б)? — оба дают состоятельные оценки β.

(г) Для CHOICE Data найдите оценки коэффициентов уравнения $P(school_i=1)=\Lambda(x'\beta)$, где School =1 если at16=1 и 0 иначе, а x={able7, loginc, ctratio, oldsib,yngsib, etot, female}.
Приведите код STATA и сравните полученные оценки.

Какие оценки дают больше точных совпадений при прогнозе $\hat{y}_i=1$, если $\hat{p}_i=\Lambda(x_i'\hat{\beta})>0.5$?

Описание данных:

CHOICE Data. In the UK, an important career choice is made at age 16. At this age, all children sit national exams. A few months later, they have to decide whether to stay at school or to leave full-time education. If they leave, they can choose between a regular job, or some type of apprenticeship, combining education with work. In this set of exercises, we will examine which factors determine this choice.

We will use data from the UK National Child and Development Survey. This data set covers individuals born in the UK in March 1958. See Micklewright (1986) for a detailed description of this data source. Data on these respondents are collected at various stages of their life cycle. We use a subsample of boys and girls, excluding those living in Scotland. Most of the variables we use are measured at age 16.

At16	Continuation decision taken at age 16: 1: stays at school, 2: apprenticeship, 3: regular job.
Able7	General ability test score, measured at age 7.
Loginc	Log family income (at age 16)
Ctratio	Number of children per teacher at school level (school quality indicator)
Oldsib	Number of older siblings (at age 16)
Yngsib	Number of younger siblings (at age 16)
Etot	Number of O-levels obtained at national exams at age 16 (prior to continuation decision)
Female	1 for girls, 0 for boys

Solution

(а)

$$\frac{\partial \ln L}{\partial\beta}=\sum\left(y_i\frac{\lambda(x_i'\beta)}{\Lambda(x_i'\beta)}x_i-(1-y_i)\frac{\lambda(x_i'\beta)}{1-\Lambda(x_i'\beta)}x_i\right)=\sum\left(y_i\left(\frac{\lambda(x_i'\beta)}{\Lambda(x_i'\beta)}+\frac{\lambda(x_i'\beta)}{1-\Lambda(x_i'\beta)}\right)-\frac{\lambda(x_i'\beta)}{1-\Lambda(x_i'\beta)}\right)x_i=$$

$$=\sum\left(y_i-\frac{\lambda(x_i'\beta)}{1-\Lambda(x_i'\beta)}\right)x_i=\sum\left(y_i-\Lambda(x_i'\beta)\right)x_i=0$$

(б) $\frac{\partial g(\beta)}{\partial\beta}=-2\sum_{i=1}^{n}\left(y_i-\Lambda(x_i'\beta)\right)\lambda(x_i'\beta)x_i=0$.

(в) Не совпадают.

(г)

```
gen school = (at16==1)
```

logit school able7 loginc ctratio oldsib yngsib etot female

```
Logistic regression                                     Number of obs =  3,423
                                                        LR chi2(7)    = 964.33
                                                        Prob > chi2   = 0.0000
Log likelihood = -1645.6347                             Pseudo R2     = 0.2266

------------------------------------------------------------------------------
      school | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       able7 |   .0407739   .0031581    12.91   0.000     .0345842    .0469636
      loginc |   .6760199     .11748     5.75   0.000     .4457633    .9062765
     ctratio |  -.2428075   .0241752   -10.04   0.000    -.2901901   -.1954249
      oldsib |  -.2935737   .0743979    -3.95   0.000    -.4393908   -.1477566
      yngsib |  -.0634156   .0379222    -1.67   0.094    -.1377418    .0109106
        etot |    .221943   .0170205    13.04   0.000     .1885835    .2553026
      female |  -.0590708   .0858011    -0.69   0.491    -.2272378    .1090963
       _cons |  -3.795431   .6840002    -5.55   0.000    -5.136047   -2.454816
------------------------------------------------------------------------------
```

predict plogit, pr
gen ylog =(plogit>0.5)
tab school ylog

```
           |         ylog
    school |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,068        283 |     2,351
         1 |       502        570 |     1,072
-----------+----------------------+----------
     Total |     2,570        853 |     3,423
```

nl (school = logistic({b0} +{b1}*able7 +{b2}*loginc ///
+{b3}*ctratio +{b4}*oldsib +{b5}*yngsib +{b6}*etot +{b7}*female))

```
      Source |       SS           df       MS
-------------+----------------------------------    Number of obs =      3,423
       Model |  531.08309          8  66.3853859    R-squared     =     0.4954
    Residual |  540.91691       3415  .158394411    Adj R-squared =     0.4942
-------------+----------------------------------    Root MSE      =    .397988
       Total |       1072       3423  .313175577    Res. dev.     =   3398.594

------------------------------------------------------------------------------
      school | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         /b0 |  -4.441313   .6728116    -6.60   0.000    -5.760467   -3.122159
         /b1 |   .0462135   .0036355    12.71   0.000     .0390855    .0533416
         /b2 |   .6657039   .1091681     6.10   0.000     .4516625    .8797453
         /b3 |  -.2502805   .0235902   -10.61   0.000    -.2965329   -.2040282
         /b4 |  -.2701057   .0692833    -3.90   0.000    -.4059466   -.1342647
         /b5 |    -.04579   .0354875    -1.29   0.197    -.1153688    .0237888
         /b6 |   .2626527   .0190402    13.79   0.000     .2253213    .2999841
         /b7 |  -.0664962   .0781859    -0.85   0.395     -.219792    .0867997
------------------------------------------------------------------------------
```

predict pnl, yhat
gen ynl =(pnl>0.5)
tab school ynl

```
           |          ynl
    school |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,065        286 |     2,351
         1 |       494        578 |     1,072
-----------+----------------------+----------
     Total |     2,559        864 |     3,423
```

Больше при NL: 578 > 570/

Сравним коэффициенты:

	NL	Logit
able7	0.0462	0.0408
loginc	0.6657	0.6760

ctratio	-0.2503	-0.2428
oldsib	-0.2701	-0.2936
yngsib	-0.0458	-0.0634
etot	0.2627	0.2219
female	-0.0665	-0.0591
_cons	-4.4413	-3.7954

Problem 2

Let you have the model $y_{it} = x_{it}'\beta + c_i + u_{it},\ i = 1,...,n;\ t = 1,2.$

Show that FE and FD estimators are numerically identical.

Solution

FE estimator is OLS estimator in the equation $y_{i2} - \bar{y}_i = (x_{i2} - \bar{x}_i)'\beta + (u_{i2} - \bar{u}_t),$

$y_{i2} - \bar{y}_i = y_{i2} - \frac{1}{2}(y_{i1} + y_{i2}) = \frac{1}{2}(y_{i2} - y_{i1}) = \frac{1}{2}\Delta y_{i2}$, same for x and *u*. Thus this equation is

$\frac{1}{2}\Delta y_{i2} = \frac{1}{2}(\Delta x_{i2})'\beta + \frac{1}{2}\Delta u_{i2}$, or $\Delta y_{i2} = (\Delta x_{i2})'\beta + \Delta u_{i2}$ — just the equation for FD estimator.

Problem 3

The data are taken from the National Longitudinal Survey (NLS Youth Sample) and contain observations on 545 males for the years 1980–1987. The first two variables indicate the NLS individual identification number and the year of observation.

The meaning of the variables and their sample means are as follows.

Variable	Definition	Mean	Standard deviation
NR	Observations number		
YEAR	Year of observation		
school	Years of schooling	11.76	1.75
exper	Age-6-School	6.51	2.83
exper2	Experience Squared	50.42	40.78
union	Wage set by collective bargaining	.24	.43
mar	Married	.44	.50
health	Has health disability	.02	.13
rural	Lives in rural area	.20	.40
wage	Log of hourly wage	1.65	.53

Use computer outputs below to answer the questions:

(a) Please explain STATA commands 1)–8) below

(b) Interpret all tests, which compare the three models. Which model do you prefer according to these tests?

(c) Why is `school` omitted in model 3?

(d) Interpret coefficients at `exper, exper2, union, health` in model 3.

(e) What is a potential problem with all these models? How would you suggest solving it?

1). xtset NR YEAR

```
panel variable:  NR (strongly balanced)
time variable:  YEAR, 1980 to 1987
delta:  1 unit
```

2). reg wage union school exper exper2 health rural mar

```
Source       SS       df       MS              Number of obs =    4360
                                                F(  7,  4352) =  145.18
Model     234.082729     7   33.4403898         Prob > F      =  0.0000
Residual  1002.44691  4352  .230341662         R-squared     =  0.1893
                                                Adj R-squared =  0.1880
Total     1236.52964  4359  .283672779         Root MSE      =  .47994

wage        Coef.   Std. Err.      t       P>t
union      .1665411   .0169746     9.81   0.000
school      .095026   .0046261    20.54   0.000
exper      .0843186   .0100833     8.36   0.000
exper2    -.0026135   .0007058    -3.70   0.000
health    -.0462718   .0563476    -0.82   0.412
rural     -.1360885   .0184257    -7.39   0.000
mar        .1409735   .0156719     9.00   0.000
```

```
_cons     .0394608   .0636711     0.62   0.535
```

3). xtreg wage union school exper exper2 health rural mar, re

```
Random-effects GLS regression                   Number of obs      =      4360
Group variable: NR                              Number of groups   =       545
R-sq:  within  = 0.1767                         Obs per group: min =         8
between = 0.1758                                               avg =       8.0
overall = 0.1762                                               max =         8

                                                Wald chi2(7)       =    932.94
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

wage       Coef.     Std. Err.      z    P>z
union     .1046705   .0178157     5.88   0.000
school     .100522   .0086879    11.57   0.000
exper     .1108428   .0082865    13.38   0.000
exper2   -.0040037   .0005938    -6.74   0.000
health   -.0222598   .0465001    -0.48   0.632
rural    -.0232938   .0239292    -0.97   0.330
mar       .0680642   .0167559     4.06   0.000
_cons    -.1042096   .1067248    -0.98   0.329
sigma_u    .3209816
sigma_e    .3512087
rho   .45512266   (fraction of variance due to u_i)
```

4). xttest0

```
Breusch and Pagan Lagrangian multiplier test for random effects

wage[NR,t] = Xb + u[NR] + e[NR,t]

Estimated results:
Var     sd = sqrt(Var)
---------+-----------------------------
wage    .2836728       .5326094
e       .1233476       .3512087
u       .1030292       .3209816

Test:        Var(u) = 0
chibar2(01) =  3101.01
Prob > chibar2 =   0.0000
```

5). est store RAN

6). xtreg wage union school exper exper2 health rural mar, fe

```
note: school omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      4360
Group variable: NR                              Number of groups   =       545

R-sq:  within  = 0.1787                         Obs per group: min =         8
between = 0.0001                                               avg =       8.0
overall = 0.0567                                               max =         8
                                                F(6,3809)          =    138.12
corr(u_i, Xb)  = -0.1381                        Prob > F           =    0.0000

wage       Coef.     Std. Err.      t      P>t
union     .0813746   .0192956     4.22   0.000
school    (omitted)
exper     .1177057   .0084348    13.95   0.000
exper2   -.0043707   .0006066    -7.20   0.000
health   -.0169087   .0471913    -0.36   0.720
rural      .049214   .0290048     1.70   0.090
mar       .0451466   .0183138     2.47   0.014
_cons     1.053299   .0276307    38.12   0.000

     sigma_u    .40393965
     sigma_e    .3512087
         rho    .56948976   (fraction of variance due to u_i)

F test that all u_i=0:     F(544, 3809) =     7.94           Prob > F = 0.0000
```

7). est store FIX

8). hausman FIX RAN

```
---- Coefficients ----
             (b)          (B)            (b-B)       sqrt(diag(V_b-V_B))
             FIX          RAN         Difference          S.E.

union      .0813746     .1046705       -.0232959        .0074107
exper      .1177057     .1108428        .0068629        .0015746
exper2    -.0043707    -.0040037        -.000367         .000124
health    -.0169087    -.0222598         .005351        .0080471
rural       .049214    -.0232938        .0725078        .0163912
mar        .0451466     .0680642       -.0229176        .0073914

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test:  Ho:  difference in coefficients not systematic

chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
=       59.83
Prob>chi2 =      0.0000
```

Solution

(a) (6 points) 1) set panel data structure; 2), 3) and 6) — pooled, RE and FE models; 4) test to discriminate between RE–pooled; 5),7) — store estimates data for the Hausman test 8).

(b) (6 points) Breuch-Pagan test reject pooled for RE; F-test reject pooled for FE (F(544, 3809)=7.94); Hausman test reject RE for FE. So we prefer FE model.

(c) (6 points) school is not changed over time and is included in fixed effect.

(d) (6 points) union membersheep increase wage by 8%, wage is increased with experience, but marginal effect decreases. “optimal” exper is approx. 13.5 years; health is not significant.

(e) (6 points) Endogeneity. The problem could be with union variable. (and with mar also :))). May consider to use IV estimator.