Seminar 3 -- Bootstrap

You plan to study the marriage behavior for the US Black women in the Midwest region. The data you will use is a sample from the March 2009 US Current Polulation Survey. The data file and description can be downloaded from Canvas. We use the model

$$\Pr(married = 1|x_i) = \Phi(x_i'\beta)$$

where x_i is a vector of observable charateristics for individual i, including age, age square and education. That is

$$x_i'\beta = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ.$$

1. Read the data description file and construct the sample. You should have 433 observations in you data. Define a binary variable "mar" as 1 if married and 0 otherwise from the variable "marital"

```
import stata_setup
stata_setup.config("C:/Program Files/Stata17", "se")
```

```
  ___  ____  ____  ____  ____ ©
 /__    /   ____/   /   ____/      17.0
___/   /   /___/   /   /___/       SE—Standard Edition

 Statistics and Data Science       Copyright 1985-2021 StataCorp LLC
                                   StataCorp
                                   4905 Lakeway Drive
                                   College Station, Texas 77845 USA
                                   800-STATA-PC        https://www.stata.com
                                   979-696-4600        stata@stata.com

Stata license: Unlimited-user network, expiring 21 Mar 2022
Serial number: 401709320653
  Licensed to: zhiyang Jia
               UIO

Notes:
      1. Unicode is supported; see help unicode_advice.
      2. Maximum number of variables is set to 5,000; see help set_maxvar.
```

```
%%stata

clear all

set seed 2345688

use cps09mar


su

keep if region==2 & race==2 & female==1

gen mar=(marital<4)
```

```
.
. clear all

.
. set seed 2345688

.
. use cps09mar
(Written by R.              )

.
.
. su

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         age |     50,742    42.13173    11.48762         15         85
      female |     50,742    .4257223    .4944569          0          1
        hisp |     50,742    .1487919    .3558868          0          1
   education |     50,742    13.92462    2.744447          0         20
    earnings |     50,742    55091.53    52222.07          1     561087
-------------+---------------------------------------------------------
       hours |     50,742    43.82724    7.704467         36         99
        week |     50,742    51.87927    .5986461         48         52
       union |     50,742    .0215206    .1451134          0          1
       uncov |     50,742    .0022072    .0469299          0          1
      region |     50,742    2.635627    1.060051          1          4
-------------+---------------------------------------------------------
        race |     50,742    1.433507     1.31743          1         21
     marital |     50,742    2.763174    2.503158          1          7

.
. keep if region==2 & race==2 & female==1
(50,309 observations deleted)

.
. gen mar=(marital<4)

.
```

2. Specify the ML function and estimate the probit model.

Answer hint

$$\begin{aligned} L_N(\beta) &= \sum y_i log(P_i) + (1-y_i) log(1-P_i) \\ &= \sum y_i log(\Phi(x_i\beta)) + (1-y_i) log(1-\Phi(x_i\beta)) \end{aligned}$$

```
%%stata
gen age2=age*age

probit mar age age2 education, r

est store probit_ml
```

```
. gen age2=age*age

.
. probit mar age age2 education, r

Iteration 0:   log pseudolikelihood = -285.21583
Iteration 1:   log pseudolikelihood = -272.00107
Iteration 2:   log pseudolikelihood = -271.96692
Iteration 3:   log pseudolikelihood = -271.96692

Probit regression                                       Number of obs =    433
                                                        Wald chi2(3)  =  26.03
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -271.96692                       Pseudo R2     = 0.0465

------------------------------------------------------------------------------
             |               Robust
         mar | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .1380565   .0507529     2.72   0.007     .0385825    .2375304
        age2 |  -.0014209   .0006022    -2.36   0.018    -.0026013   -.0002406
   education |   .0898554   .0272241     3.30   0.001     .0364971    .1432137
       _cons |  -4.714038   1.103265    -4.27   0.000    -6.876397   -2.551678
------------------------------------------------------------------------------

.
. est store probit_ml

.
```

3. Explain how you can calculate the bootstrap standard error

Answer hint

1. Computes the coefficient estimates $\hat{\beta}$ on the estimation sample.
2. Draws n observations at random (with replacement) from the estimation sample
3. Computes the estimates $\hat{\beta}^{(*r)}$ on this simulated sample
4. Repeats this R times, obtaining

$$\widehat{se}^*(\hat{\beta}) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\beta}^{(*r)} - \overline{\hat{\beta}^{(*)}}\right)^2}$$

$\widehat{se}^*(\hat{\beta}) = 0.0287$

```
%%stata


bootstrap "probit mar age age2 education, r" _b _se, ///
   reps(1000) saving(probit_bsdata) replace

preserve

use probit_bsdata, clear

su b_education
```

```
.
.
. bootstrap "probit mar age age2 education, r" _b _se, ///
>    reps(1000) saving(probit_bsdata) replace

Command:      probit mar age age2 education , r
Statistics:   b_age      = [mar]_b[age]
              b_age2     = [mar]_b[age2]
              b_educat~n = [mar]_b[education]
              b_cons     = [mar]_b[_cons]
              se_age     = [mar]_se[age]
              se_age2    = [mar]_se[age2]
              se_educa~n = [mar]_se[education]
              se_cons    = [mar]_se[_cons]

Bootstrap statistics                            Number of obs    =        433
                                                Replications     =       1000

------------------------------------------------------------------------------
Variable      |  Reps  Observed      Bias  Std. err. [95% conf. interval]
--------------+---------------------------------------------------------------
        b_age |  1000  .1380565  .0089082  .0537027   .0326735   .2434394   (N)
              |                                       .0565371   .2669305   (P)
              |                                       .0526944   .2544489  (BC)
       b_age2 |  1000 -.0014209 -.0001033  .0006393  -.0026754  -.0001664   (N)
              |                                      -.0029622  -.0004738   (P)
              |                                      -.0028189  -.0004084  (BC)
  b_education |  1000  .0898554  .0019074  .0286487   .0336369   .1460739   (N)
              |                                       .0357806    .148678   (P)
              |                                       .0304013   .1428906  (BC)
       b_cons |  1000 -4.714037 -.2134345   1.15536  -6.981248  -2.446827   (N)
              |                                      -7.499677  -2.978604   (P)
              |                                      -7.105587   -2.82808  (BC)
       se_age |  1000  .0507529 -.0024728  .0064473   .0381012   .0634047   (N)
              |                                       .0369542   .0632855   (P)
              |                                       .0434362   .0758202  (BC)
      se_age2 |  1000  .0006022 -.0000322  .0000808   .0004437   .0007607   (N)
              |                                       .0004273   .0007516   (P)
              |                                        .000509   .0009107  (BC)
 se_education |  1000  .0272241  .0004513   .001517   .0242472    .030201   (N)
              |                                        .024462   .0304959   (P)
              |                                       .0235468   .0295213  (BC)
      se_cons |  1000  1.103265 -.0386056  .1277354   .8526046   1.353926   (N)
              |                                       .8436219   1.356709   (P)
              |                                       .9489852   1.634393  (BC)
------------------------------------------------------------------------------
Key:  N: Normal
      P: Percentile
     BC: Bias-corrected

.
. preserve

.
. use probit_bsdata, clear
(bootstrap: probit mar age age2 education , r)

.
. su b_education

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
 b_education |      1,000    .0917628    .0286487   .0025177   .1830203

.
.
```

4. Now you want to test the hypothesis that the level of education does not have any effect on the marriage decision. You are asked to use different methods:

(a) t-test using both ML and Bootstrapped standard errors.

Answer hint

$$t_{ml} = \frac{0.0898 - 0}{0.0272} = 3.30$$

$$t_{boot} = \frac{0.0898 - 0}{0.0287} = 3.13$$

(b) Confidence interval method

1. using the percentile method

Answer hint

Using the bootstrapped standard errors we can the percentile method, which gives

$$[\hat{\beta}^*_{\alpha/2}, \hat{\beta}^*_{1-\alpha/2}]$$

where $\hat{\beta}^*_p$ is the p-th quantile of the bootstrap distribution of $\hat{\beta}^{(*)}$

```
%%stata

* kdensity b_education,normal

_pctile b_education, p(2.5, 97.5)
* display quantiles
dis r(r2)
dis r(r1)
```

```
.
. * kdensity b_education,normal
.
. _pctile b_education, p(2.5, 97.5)

. * display quantiles
. dis r(r2)
.14867803

. dis r(r1)
.03578065

.
```

1. the percentile t-method

Construct for each bootstrap sample

$$t^{(*r)} = \frac{\hat{\beta}^{(*r)} - \hat{\beta}}{\widehat{se}^{(*r)}(\hat{\beta})}$$

The percentile t-method defines interval

$$[\hat{\beta} - t^*_{1-\alpha/2}\hat{\sigma}, \hat{\beta} - t^*_{\alpha/2}\hat{\sigma}]$$

where t^*_p is the p-th quantile of the bootstrap distribution of t^*.

```
%%stata

est restore probit_ml

* create bootstrap t-ratios
gen t_edu=(b_education-_b[education])/se_education
* calculate qunaitles of t-ratios
_pctile t_edu, p(2.5, 97.5)
* display quantiles
dis r(r2)
dis r(r1)

* the confidence interval
dis _b[education]-_se[education]*r(r2)
dis _b[education]-_se[education]*r(r1)
```

```
.
. est restore probit_ml
(results probit_ml are active now)

.
. * create bootstrap t-ratios
. gen t_edu=(b_education-_b[education])/se_education

. * calculate qunaitles of t-ratios
. _pctile t_edu, p(2.5, 97.5)

. * display quantiles
. dis r(r2)
2.0045297

. dis r(r1)
-2.0948982

.
. * the confidence interval
. dis _b[education]-_se[education]*r(r2)
.03528388

. dis _b[education]-_se[education]*r(r1)
.14688712

.
```

(c) What the difference between these two different confidence interval? Which one you would prefer and why?

The first

- is very simple to implement.
- works well if the underlying distribution can be transformed to a normal.
- no theoretical evidence on whether it will be better than the standard CI from asymptotics.

The second

- Provide asymptotic refinement (are second order accurate) regardless of the true distribution

5. What will be the bias corrected estimate for the parameter for age?

answer hint

The bias-corrected estimate of β_1 equals

$$\hat{\beta}_1 - \text{Bias}^* = \hat{\beta}_1 - \left(\overline{\hat{\beta}_1^{(*)}} - \hat{\beta}_1\right) = 2\hat{\beta}_1 - \overline{\hat{\beta}_1^{(*)}}$$

```
%%stata

su b_education
dis 2*_b[education]-r(mean)
```

```
.
. su b_education

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
 b_education |      1,000    .0917628    .0286487   .0025177   .1830203

. dis 2*_b[education]-r(mean)
.08794801

.
```

6. We want to estimate at what age women are most likely to be married.

(a) Can you propose a estimator for this purpose? Implement your estimator using our data.

answer hint

The age at which the probability is maximized (if $\beta_1 > 0$ and $\beta_1 < 0$) is $\theta = \frac{-\beta_1}{2\beta_2}$.

This is easily estimated by $\hat{\theta} = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$.

(b) How can you obtain standard error for your proposed estimator? (hint: the Delta method and the Bootstrap method).

answer hint

$$Var\left(\frac{X}{Y}\right) \approx \left(\frac{1}{\mu_Y}\right)^2 \left[\sigma_X^2 + \left(\frac{\mu_X}{\mu_Y}\right)^2 \sigma_Y^2 - 2\left(\frac{\mu_X}{\mu_Y}\right)\sigma_{X,Y}\right]$$

Because of the nonlinearity, bootstrap standard errors might be more reliable.

(c) Test the hypothesis that this age is 40.

```
%%stata

nlcom (theta: (-_b[age]/(_b[age2]*2)))

testnl (-_b[age]/(_b[age2]*2))=40

capture erase theta_bsdata.dta

restore

bootstrap theta=(-_b[age]/(_b[age2]*2)), reps (100) ///
   saving(theta_bsdata): probit mar age age2 education,r

use theta_bsdata, clear

su theta


_pctile theta, p(2.5, 97.5)
* display quantiles
dis r(r2)
dis r(r1)
```

```
.
. nlcom (theta: (-_b[age]/(_b[age2]*2)))

       theta: (-_b[age]/(_b[age2]*2))

------------------------------------------------------------------------------
         mar | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       theta |   48.57935   3.524436    13.78   0.000     41.67158    55.48712
------------------------------------------------------------------------------

.
. testnl (-_b[age]/(_b[age2]*2))=40

  (1)  (-_b[age]/(_b[age2]*2)) = 40

               chi2(1) =        5.93
           Prob > chi2 =        0.0149

.
. capture erase theta_bsdata.dta

.
. restore

.
. bootstrap theta=(-_b[age]/(_b[age2]*2)), reps (100) ///
>    saving(theta_bsdata): probit mar age age2 education,r
(running probit on estimation sample)

Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

Probit regression                                     Number of obs = 433
                                                      Replications  = 100

      Command: probit mar age age2 education, r
        theta: -_b[age]/(_b[age2]*2)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       theta |   48.57935   3.554243    13.67   0.000     41.61316    55.54554
------------------------------------------------------------------------------

.
. use theta_bsdata, clear
(bootstrap: probit)

.
. su theta

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       theta |        100    48.68006    3.554243     42.423   64.67423

.
.
. _pctile theta, p(2.5, 97.5)

. * display quantiles
. dis r(r2)
56.954414

. dis r(r1)
43.852783

.
.
```

(d) Can you present a graphic representation of the distribution of your proposed estimator?

answer hint

We can use the bootstrapped distribution $\hat{F}$ to approximate the distribution F. We generate a kernel density plot of the bootstrapped estimator.

```
%%stata

kdensity theta
```

```
.
. kdensity theta

.
```