Probit Regression in R, Python, Stata, and SAS

Shi Lan, Roya Talibova, Bo Qu,Jiehui Ding

2018/11/26

- Model Introduction
- Dataset: Mroz
- Languages

Model Introduction

(tab content)

Dataset: Mroz

(tab content)

Languages

R Python Stata SAS

1.Data Summary

Firstly, We import the Mroz data from website and show the first six rows of the dataset.

```
*Importing data
import delimited https://vincentarelbundock.github.io/Rdatasets/csv/carData/Mroz.csv,
clear
save mroz,replace
use mroz,clear
*List the first six rows
list if v1<=6
```

	v1	lfp	k5	k618	age	wc	hc	lwg	inc
1.	1	yes	1	0	32	no	no	1.210165	10.91
2.	2	yes	0	2	30	no	no	.3285041	19.5
3.	3	yes	1	3	35	no	no	1.514128	12.04
4.	4	yes	0	3	34	no	no	.0921151	6.8
5.	5	yes	1	2	31	yes	no	1.52428	20.1
6.	6	yes	0	0	54	no	no	1.556486	9.859

Then, We change all binary variables to be numeric, and we get a summary of the data. Our response is lfp and its mean is 0.57. The range of age is from 30 to 60.

```
*Change variables with values yes/no to 1/0
gen lfpart =1 if lfp == "yes"
replace lfpart =0 if lfp == "no"
gen wifec =1 if wc == "yes"
replace wifec =0 if wc == "no"
gen husbc =1 if hc == "yes"
replace husbc =0 if hc == "no"
drop lfp wc hc
rename lfpart lfp
rename wifec wc
rename husbc hc
*Get the summary of the data
summ
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          v1 |        753         377    217.5167          1        753
          k5 |        753    .2377158     .523959          0          3
        k618 |        753    1.353254    1.319874          0          8
         age |        753    42.53785    8.072574         30         60
         lwg |        753    1.097115    .5875564  -2.054124   3.218876
-------------+---------------------------------------------------------
         inc |        753    20.12897     11.6348      -.029         96
         lfp |        753    .5683931    .4956295          0          1
          wc |        753    .2815405    .4500494          0          1
          hc |        753    .3917663    .4884694          0          1
```

2.Fitting model by Probit Regression

Now, we fit our data by probit regression. lfp is the response and the remaining variables are predictors. Looking at the p-values, all variables have highly significant, except k618 and hc.

```
*Fitting the data by probit regression
probit lfp k5 k618 age lwg inc i.wc i.hc
```

```
Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -452.84838
Iteration 2:   log likelihood = -452.69498
Iteration 3:   log likelihood = -452.69496

Probit regression                               Number of obs     =         75
> 3
                                                LR chi2(7)        =      124.3
> 6
                                                Prob > chi2       =      0.000
> 0
Log likelihood = -452.69496                     Pseudo R2         =      0.120
> 8

------------------------------------------------------------------------------
> -
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval
> ]
-------------+----------------------------------------------------------------
> -
          k5 |  -.8747111   .1135584    -7.70   0.000    -1.097281   -.652140
> 8
        k618 |  -.0385945   .0404893    -0.95   0.340    -.1179521    .040763
> 1
         age |  -.0378235   .0076093    -4.97   0.000    -.0527375   -.022909
> 5
         lwg |   .3656287   .0877792     4.17   0.000     .1935846    .537672
> 7
         inc |   -.020525   .0047769    -4.30   0.000    -.0298875   -.011162
> 5
        1.wc |   .4883144   .1354873     3.60   0.000     .2227641    .753864
> 7
        1.hc |   .0571703   .1240053     0.46   0.645    -.1858755    .300216
> 2
       _cons |   1.918422   .3806539     5.04   0.000     1.172354     2.6644
> 9
------------------------------------------------------------------------------
> -
```

We get a summary of the probit prediction from the fitted model, we get the smallest probability is 0.005691 and the largest probability is 0.9745. The 50% percentile is 0.5782336, which is close to its mean we showed above.

```
*Predicting the probability of labor-force  participation
predict prob_lfp
summ prob_lfp, detail
```

```
                           Pr(lfp)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0874537        .005691
 5%     .2087887       .0280799
10%     .3134367       .0322375       Obs                 753
25%     .4470239        .056195       Sum of Wgt.         753

50%     .5782336                      Mean           .5705144
                        Largest       Std. Dev.      .1928416
75%     .7189098       .9530371
90%     .8133735       .9554808       Variance       .0371879
95%     .8603116        .966253       Skewness      -.3429077
99%     .9348801       .9744748       Kurtosis       2.709472
```

3.Marginal effect

Now, we predict the data for groups defined by levels of categorical variables. ##### Group by hc First, we make a table of frequently count of hc and lfp we predict the lfp for two groups: hc=0 and hc=1, and we keep other variables at mean.

```
tab lfp hc
```

```
           |          hc
       lfp |         0          1 |     Total
-----------+----------------------+----------
         0 |       207        118 |       325
         1 |       251        177 |       428
-----------+----------------------+----------
     Total |       458        295 |       753
```

```
*use margins for each level of hc
margins hc, atmeans
```

```
Adjusted predictions                            Number of obs     =        75
> 3
Model VCE    : OIM

Expression   : Pr(lfp), predict()
at           : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =    42.53785 (mean)
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)

------------------------------------------------------------------------------
> -
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval
> ]
-------------+----------------------------------------------------------------
> -
          hc |
          0  |   .5693818   .0273369    20.83   0.000     .5158024    .622961
> 1
          1  |   .5917197   .0345427    17.13   0.000     .5240172    .659422
> 1
------------------------------------------------------------------------------
> -
```

The marginal probability of hc=1 (husband has attained college) is 0.59 and it slightly higher than the marginal probability of hc=0 (husband has not attained college), which is 0.57. There is not obivious differnce. It is reasonable because the p-value of hc is very high.

Group by wc

The table of frequently shows that when wc=0, the proportion of lfp is average, which is closed to 0.5. However, when wc=1, the proportion of lfp=1 is much higher.

```
tab lfp wc
```

```
            |          wc
        lfp |         0          1 |     Total
------------+----------------------+----------
          0 |       257         68 |       325
          1 |       284        144 |       428
------------+----------------------+----------
      Total |       541        212 |       753
```

we predict the lfp for two groups: wc=0 and wc=1, and we keep other variables at mean.

```
*use margins for each level of wc
margins wc, atmeans
```

```
Adjusted predictions                              Number of obs     =        75
> 3
Model VCE    : OIM

Expression   : Pr(lfp), predict()
at           : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =    42.53785 (mean)
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)

------------------------------------------------------------------------------
> -
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval
> ]
-------------+----------------------------------------------------------------
> -
          wc |
          0  |   .5238097   .0241197    21.72   0.000     .4765359     .571083
> 6
          1  |    .708165   .0380449    18.61   0.000     .6335984     .782731
> 6
------------------------------------------------------------------------------
> -
```

The result shows that the marginal probability is 0.71 when wc=1 and the marginal probability is 0.52 when wc=0. The probability of participating labor-force is higher when wife has attended college. We can say that wife's college attendance is an important predictor.

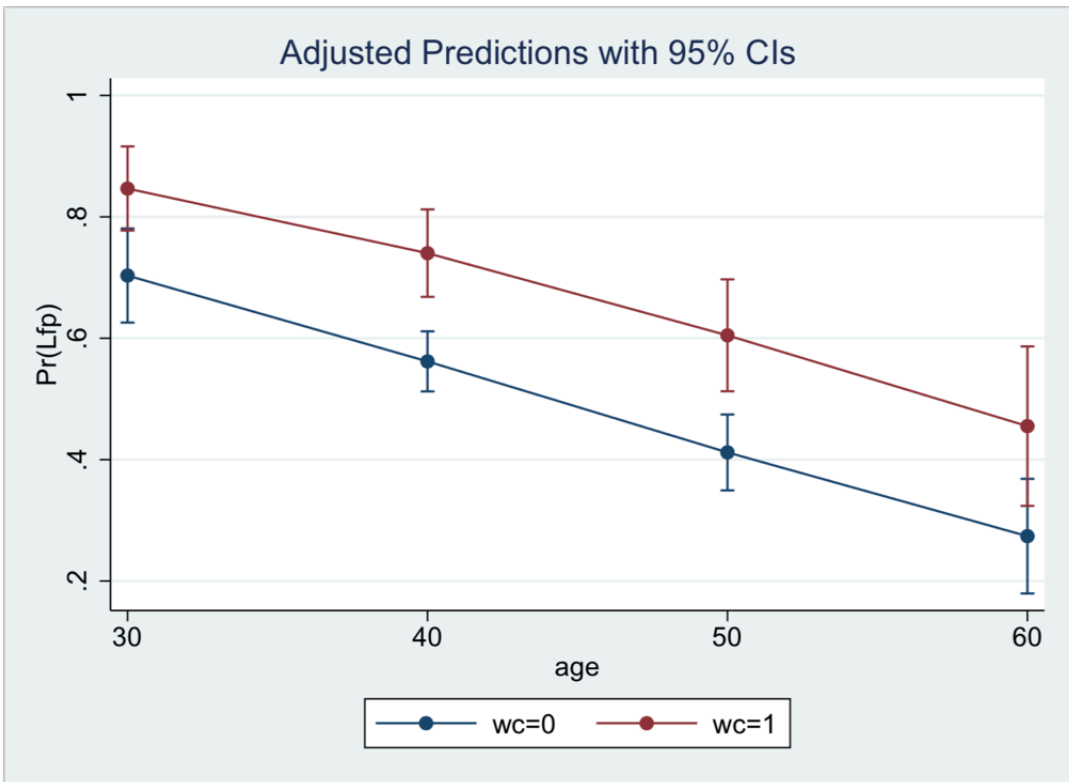
We can go deeper on the predictor wc. We predict lfp for group by age and wc. Age is at every 10 years of age from 30 to 60. Since the output of marginal function is long, we make a plot to visualize the output and it is easier to interpert.

```
*use margins for each level of wc and age
margins, at(age=(30(10)60) wc=(0 1)) atmeans vsquish
```

```
Adjusted predictions                          Number of obs     =         75
> 3
Model VCE    : OIM

Expression   : Pr(lfp), predict()
1._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          30
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           0
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
2._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          30
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           1
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
3._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          40
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           0
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
4._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          40
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           1
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
5._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          50
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           0
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
6._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          50
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           1
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
7._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          60
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           0
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
8._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               age             =          60
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               wc              =           1
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
----------------------------------------------------------------------------
> -
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval
> ]
-------------+--------------------------------------------------------------
```

```
> -
        _at |
          1 |   .7033095   .0395332    17.79   0.000     .6258258    .780793
> 2
          2 |   .8466704   .0353618    23.94   0.000     .7773626    .915978
> 3
          3 |   .5618684   .0252363    22.26   0.000     .5124062    .611330
> 6
          4 |   .7402195   .0367492    20.14   0.000     .6681924    .812246
> 6
          5 |   .4119518   .0319053    12.91   0.000     .3494185    .474485
> 1
          6 |   .6047985   .0470611    12.85   0.000     .5125605    .697036
> 5
          7 |   .2739992    .048177     5.69   0.000     .1795741    .368424
> 4
          8 |   .4552342   .0670442     6.79   0.000       .32383    .586638
> 4
------------------------------------------------------------------------------
> -
```

```
marginsplot
```



From the marginal plot, we can conclude that when age is increasing, the probability is decreasing. Also, The probability of wc=1 is always higher than wx=0. At age 60, the variablity is the highest because the 95% confidence interval is the widest.

Group by k5

The table of frequently shows that the proportion of lfp is decreasing when k5 is increasing.

```
tab lfp k5
```

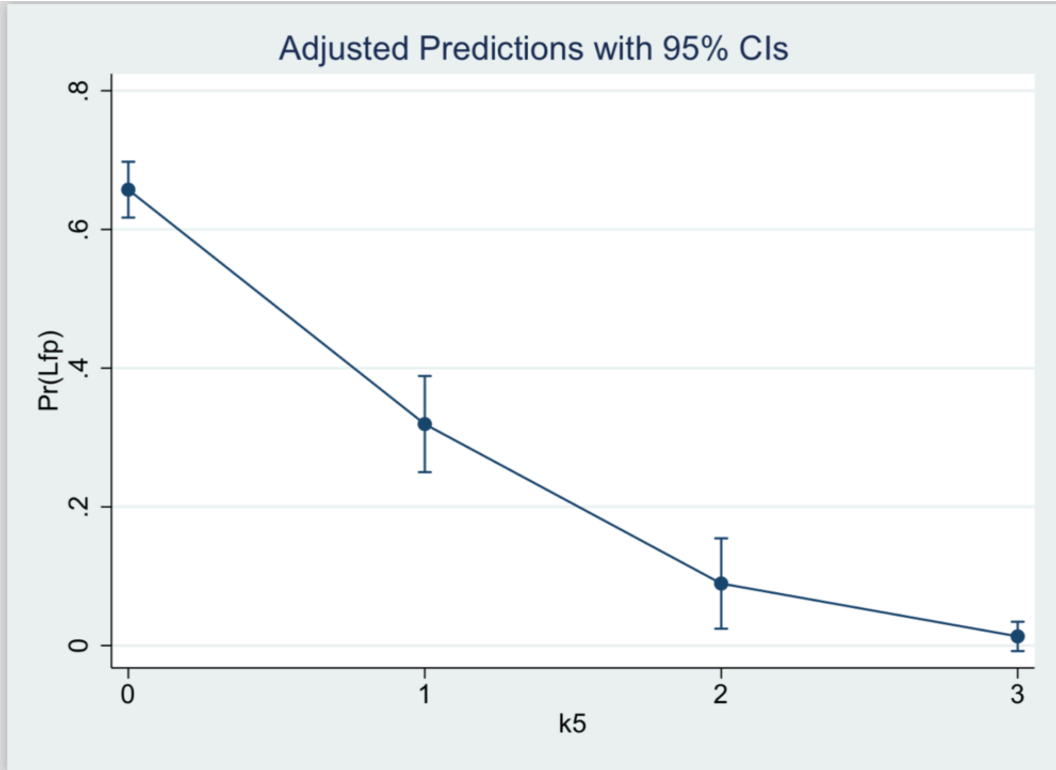
lfp	k5 0	1	2	3	Total
0	231	72	19	3	325
1	375	46	7	0	428
Total	606	118	26	3	753

we predict the lfp by k5= 0 1 2 3, and we keep other variables at mean. Also, we make a plot to visualize the data.

```
*use margins for each level of k5
margins, at(k5=(0 1 2 3)) atmeans
```

```
Adjusted predictions                            Number of obs     =         75
> 3
Model VCE    : OIM

Expression   : Pr(lfp), predict()

1._at        : k5              =           0
               k618            =    1.353254 (mean)
               age             =    42.53785 (mean)
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)

2._at        : k5              =           1
               k618            =    1.353254 (mean)
               age             =    42.53785 (mean)
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)

3._at        : k5              =           2
               k618            =    1.353254 (mean)
               age             =    42.53785 (mean)
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)

4._at        : k5              =           3
               k618            =    1.353254 (mean)
               age             =    42.53785 (mean)
               lwg             =    1.097115 (mean)
               inc             =    20.12897 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)

------------------------------------------------------------------------------
> -
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval
> ]
-------------+----------------------------------------------------------------
> -
         _at |
          1  |   .6573092   .0205632    31.97   0.000     .6170061    .697612
> 4
          2  |   .3193274   .0353742     9.03   0.000     .2499952    .388659
> 5
          3  |    .089427   .0332266     2.69   0.007      .024304    .154550
> 1
          4  |   .0132433   .0107846     1.23   0.219    -.0078942    .034380
> 7
------------------------------------------------------------------------------
> -
```

```
marginsplot
```



The output shows that when women do not have any children 5 years old or younger, the probability of participating labor-force is 0.66 which is higher than the average. However, after they had childrens, the probability of participating labor-force is decreasing. Therefore, we can conclude that k5 is a significant predictor.