ECON 452*

Overview of *Stata 12/13* Tutorials 8 and 9

1. Three Probit Models

Three models of married women's labour force participation, where the observed binary dependent variable $\boldsymbol{inlf_i}$ is defined as follows:

$\boldsymbol{inlf_i}$ = 1 if the i-th married woman is in the employed labour force
= 0 if the i-th married woman is not in the employed labour force

Probit Model 1: Has **six explanatory variables, all *continuous***

The ***probit index function***, or ***regression function***, for **Model 1** is:

$$x_i^T\beta = \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i + \beta_6 kidslt6_i + \beta_7 kidsge6_i$$

where

$nwifeinc_i$ = non-wife family income of the i-th woman (in thousands of dollars per year);
ed_i = years of formal education of the i-th woman (in years);
exp_i = years of actual work experience of the i-th woman (in years);
age_i = age of the i-th woman (in years);
$kidslt6_i$ = number of children less than 6 years of age for the i-th woman;
$kidsge6_i$ = number of children 6 years of age or older for the i-th woman.

Probit Model 2

Has **five explanatory variables**

- **four *continuous* explanatory variables:** nwifeinc_i, ed_i, $\exp_i$, and age_i
- **one *binary* explanatory variable:** dkidslt6_i

The **probit index function**, or regression function, **for Model 2** is:

$$x_i^T\beta = \beta_0 + \beta_1\text{nwifeinc}_i + \beta_2\text{ed}_i + \beta_3\exp_i + \beta_4\exp_i^2 + \beta_5\text{age}_i + \delta_0\text{dkidslt6}_i$$

where ***dkidslt6$_i$*** is a binary explanatory variable defined as follows:

dkidslt6$_i$ = 1 if $\text{kidslt6}_i > 0$ for the i-th married woman
= 0 if $\text{kidslt6}_i = 0$ for the i-th married woman

Probit Model 3

Has **five explanatory variables** (the same ones as Model 2)

- **four *continuous* explanatory variables:** nwifeinc_i, ed_i, $\exp_i$, and age_i
- **one *binary* explanatory variable:** dkidslt6_i

The **probit index function**, or regression function, **for Model 3** is:

$$\begin{aligned} x_i^T\beta = {} & \beta_0 + \beta_1 \text{nwifeinc}_i + \beta_2 \text{ed}_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 \text{age}_i \\ & + \delta_0 \text{dkidslt6}_i + \delta_1 \text{dkidslt6}_i \text{nwifeinc}_i + \delta_2 \text{dkidslt6}_i \text{ed}_i + \delta_3 \text{dkidslt6}_i \exp_i + \delta_4 \text{dkidslt6}_i \exp_i^2 + \delta_5 \text{dkidslt6}_i \text{age}_i \end{aligned}$$

Remarks: Model 3 is the *full-interaction generalization* of Model 2: it interacts the dkidslt6_i indicator variable with all the other regressors in Model 2, and thereby permits all index function coefficients to differ between the two groups of married women distinguished by dkidslt6_i.

Probit index function for Model 3 is:

$$x_i^T\beta = \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i$$
$$+ \delta_0 dkidslt6_i + \delta_1 dkidslt6_i\, nwifeinc_i + \delta_2 dkidslt6_i\, ed_i + \delta_3 dkidslt6_i \exp_i + \delta_4 dkidslt6_i \exp_i^2 + \delta_5 dkidslt6_i\, age_i$$

- In Model 3, the probit index function for ***married women who currently have no pre-school aged children***, for whom $dkidslt6_i$ = 0, is obtained by setting $dkidslt6_i$ = 0 in the index function for Model 3:

$$\left(x_i^T\beta \,\middle|\, dkidslt6_i = 0\right) = \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i$$

- In Model 3, the probit index function for ***married women who currently have one or more pre-school aged children***, for whom $dkidslt6_i$ = 1, is obtained by setting $dkidslt6_i$ = 1 in the index function for Model 3:

$$\left(x_i^T\beta \,\middle|\, dkidslt6_i = 1\right) = \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i$$
$$+ \delta_0 1 + \delta_1 1\cdot nwifeinc_i + \delta_2 1\cdot ed_i + \delta_3 1\cdot \exp_i + \delta_4 1\cdot \exp_i^2 + \delta_5 1\cdot age_i$$

$$= \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i \; + \delta_0 + \delta_1 nwifeinc_i + \delta_2 ed_i + \delta_3 \exp_i + \delta_4 \exp_i^2 + \delta_5 age_i$$

$$= (\beta_0 + \delta_0) + (\beta_1 + \delta_1) nwifeinc_i + (\beta_2 + \delta_2) ed_i + (\beta_3 + \delta_3) \exp_i + (\beta_4 + \delta_4) \exp_i^2 + (\beta_5 + \delta_5) age_i$$

Stata 12/13 **output for *probit* and *dprobit* commands**

Use Model 2 to illustrate the meaning of several summary statistics that appear in the log file output for a **probit** or **dprobit** command in *Stata 12/13*.

The ***probit index function***, or regression function, for **Model 2** is:

$$x_i^T\beta = \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i + \delta_0 dkidslt6_i$$

where ***$dkidslt6_i$*** is a binary explanatory variable defined as follows:

$dkidslt6_i$ = 1 if $kidslt6_i > 0$ for the i-th married woman
= 0 if $kidslt6_i = 0$ for the i-th married woman

Descriptive Summary Statistics for Variables in Model 2

```
. summarize inlf nwifeinc ed exp expsq age dkidslt6

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        inlf |       753    .5683931    .4956295          0          1
    nwifeinc |       753    20.12896     11.6348  -.0290575         96
          ed |       753    12.28685    2.280246          5         17
         exp |       753    10.63081     8.06913          0         45
       expsq |       753    178.0385    249.6308          0       2025
-------------+--------------------------------------------------------
         age |       753    42.53785    8.072574         30         60
    dkidslt6 |       753    .1952191    .3966327          0          1
```

Stata Probit Estimation Commands for Model 2 -- `probit`

```
. probit inlf nwifeinc ed exp expsq age dkidslt6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -410.52123
Iteration 2:   log likelihood = -407.00272
Iteration 3:   log likelihood = -406.98832

Probit regression                                 Number of obs   =        753
                                                  LR chi2(6)      =     215.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -406.98832                       Pseudo R2       =     0.2095

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0113531   .0047493    -2.39   0.017    -.0206615   -.0020447
          ed |   .1217526   .0247398     4.92   0.000     .0732634    .1702418
         exp |   .1173689   .0185819     6.32   0.000     .0809491    .1537886
       expsq |  -.0017634   .0005991    -2.94   0.003    -.0029375   -.0005892
         age |  -.0534423   .0079364    -6.73   0.000    -.0689974   -.0378871
    dkidslt6 |  -1.022174   .1452118    -7.04   0.000    -1.306784   -.7375641
       _cons |   .4815005   .4547151     1.06   0.290    -.4097247    1.372726
------------------------------------------------------------------------------
```

Stata Probit Estimation Commands for Model 2 -- `dprobit`

```
. dprobit inlf nwifeinc ed exp expsq age dkidslt6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -410.52123
Iteration 2:   log likelihood = -407.00272
Iteration 3:   log likelihood = -406.98832

Probit regression, reporting marginal effects          Number of obs =    753
                                                       LR chi2(6)    = 215.77
                                                       Prob > chi2   = 0.0000
Log likelihood = -406.98832                            Pseudo R2     = 0.2095

------------------------------------------------------------------------------
    inlf |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
nwifeinc |  -.0044306   .0018532    -2.39   0.017    20.129  -.008063 -.000798
      ed |   .0475146   .0096489     4.92   0.000   12.2869   .028603  .066426
     exp |   .0458038   .0072682     6.32   0.000   10.6308   .031558  .060049
   expsq |  -.0006882   .0002341    -2.94   0.003   178.039  -.001147 -.000229
     age |  -.0208561   .0030943    -6.73   0.000   42.5378  -.026921 -.014791
dkidslt6*|  -.3888635     .04916    -7.04   0.000   .195219  -.485215 -.292512
---------+--------------------------------------------------------------------
  obs. P |   .5683931
 pred. P |    .583103  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Maximized Log-Likelihood Value

The **maximized log-likelihood value** for Model 2, denoted in *Stata* as **`Log likelihood`**, is the maximized value of the objective function or sample log-likelihood function corresponding to ML estimates of the probit coefficient vector $\beta = \left(\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4 \quad \beta_5 \quad \delta_0\right)^T$:

$$\ln \hat{L}_1 = \sum_{i=1}^{N} Y_i \ln \Phi\left(x_i^T\hat{\beta}\right) + \sum_{i=1}^{N}(1-Y_i)\ln\left[1-\Phi\left(x_i^T\hat{\beta}\right)\right]$$

$$\begin{aligned} = \; & \sum_{i=1}^{N} Y_i \ln \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 nwifeinc_i + \hat{\beta}_2 ed_i + \hat{\beta}_3 \exp_i + \hat{\beta}_4 \exp_i^2 + \hat{\beta}_5 age_i + \hat{\delta}_0 dkidslt6_i\right) \\ & + \sum_{i=1}^{N}(1-Y_i)\ln\left[1-\Phi\left(\hat{\beta}_0 + \hat{\beta}_1 nwifeinc_i + \hat{\beta}_2 ed_i + \hat{\beta}_3 \exp_i + \hat{\beta}_4 \exp_i^2 + \hat{\beta}_5 age_i + \hat{\delta}_0 dkidslt6_i\right)\right] \end{aligned}$$

where $\hat{\beta}_{ML} = \left(\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2 \quad \hat{\beta}_3 \quad \hat{\beta}_4 \quad \hat{\beta}_5 \quad \hat{\delta}_0\right)^T$ is the vector of ML probit coefficient estimates. $\ln \hat{L}_1$ is simply the maximized value of the objective function, the sample log-likelihood function for Model 2.

Pseudo-R^2

The **pseudo-R^2 value** for Model 2, denoted in *Stata* as **`Pseudo R2`**, is given by the expression:

$$\textbf{pseudo-}\mathbf{R^2} = 1 - \frac{\ln \hat{L}_1}{\ln \hat{L}_0}$$

where $\ln \hat{L}_0 = \sum_{i=1}^{N} Y_i \ln \Phi\left(\tilde{\beta}_0\right) + \sum_{i=1}^{N} (1 - Y_i) \ln\left[1 - \Phi\left(\tilde{\beta}_0\right)\right]$ is the **maximized log-likelihood value for the *restricted* model** implied by the null hypothesis that **all the *slope* coefficients** in Model 2 are jointly **equal to *zero*** – i.e., under the null hypothesis

H_0: $\beta_1 = 0$ *and* $\beta_2 = 0$ *and* $\beta_3 = 0$ *and* $\beta_4 = 0$ *and* $\beta_5 = 0$ *and* $\delta_0 = 0$

or

$\beta_j = 0$ for all $j = 1, \ldots, 5$ *and* $\delta_0 = 0$

The ***probit index function***, or regression function, for **Model 2** is:

$$x_i^T\beta = \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i + \delta_0 dkidslt6_i$$

LR Chi-square Statistic

The **LR chi-square statistic** for Model 2, denoted in *Stata* as **`LR chi2(6)`**, is the **likelihood ratio (LR) test statistic** for testing the ***joint* significance of the *slope* coefficients**. That is, it tests the null hypothesis H_0

H_0: $\beta_1 = 0$ *and* $\beta_2 = 0$ *and* $\beta_3 = 0$ *and* $\beta_4 = 0$ *and* $\beta_5 = 0$ *and* $\delta_0 = 0$ *or* $\beta_j = 0$ for all j = 1, …, 5 *and* $\delta_0 = 0$

against the alternative hypothesis

H_1: $\beta_1 \neq 0$ *and/or* $\beta_2 \neq 0$ *and/or* $\beta_3 \neq 0$ *and/or* $\beta_4 \neq 0$ *and/or* $\beta_5 \neq 0$ *and/or* $\delta_0 \neq 0$

or

$\beta_j \neq 0$ for j = 1, …, 5 *and/or* $\delta_0 = 0$

The ***restricted probit index function***, or regression function, for **Model 2** under the null hypothesis H_0 is:

$$x_i^T\beta = \beta_0$$

The ***unrestricted probit index function***, or regression function, for **Model 2** under the alternative hypothesis H_1 is:

$$x_i^T\beta = \beta_0 + \beta_1 nwifeinc_i + \beta_2 ed_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \beta_5 age_i + \delta_0 dkidslt6_i$$

The **LR test statistic** is equal to ***twice* the difference** between

$$\ln \hat{L}_1 = \sum_{i=1}^{N} Y_i \ln \Phi\left(x_i^T \hat{\beta}_{ML}\right) + \sum_{i=1}^{N} (1 - Y_i) \ln\left[1 - \Phi\left(x_i^T \hat{\beta}_{ML}\right)\right]$$

= the maximized log-likelihood value for the ***unrestricted*** **model** corresponding to H_1

and

$$\ln \hat{L}_0 = \sum_{i=1}^{N} Y_i \ln \Phi\left(x_i^T \tilde{\beta}_{ML}\right) + \sum_{i=1}^{N} (1 - Y_i) \ln\left[1 - \Phi\left(x_i^T \tilde{\beta}_{ML}\right)\right] = \sum_{i=1}^{N} Y_i \ln \Phi\left(\tilde{\beta}_0\right) + \sum_{i=1}^{N} (1 - Y_i) \ln\left[1 - \Phi\left(\tilde{\beta}_0\right)\right]$$

= the maximized log-likelihood value for the ***restricted*** **model** corresponding to H_0.

That is, the **LR test statistic** for a joint test of all slope coefficients in Model 2 is:

$$LR = 2\left(\ln \hat{L}_1 - \ln \hat{L}_0\right).$$

The ***null distribution of LR*** under the null hypothesis H_0 is chi-square with q degrees of freedom, i.e., $\chi^2(q) = \chi^2(6)$, since the number of restrictions q = 6 for Model 2.

$LR \sim \chi^2(q) = \chi^2(6)$ under H_0 for Model 2

z statistics

The **z-statistic** for each probit coefficient estimate is just a **t-statistic** for a **two-tail test** of the null hypothesis that the probit coefficient equals zero; i.e., a t-statistic for testing the null hypothesis

H_0: $\beta_j = 0$

against the two-sided alternative hypothesis

H_1: $\beta_j \neq 0$

The **z-statistic for H_0 versus H_1** is thus:

$$z(\hat{\beta}_j) = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \sim N(0,1) \text{ under } H_0\text{: } \beta_j = 0$$

Note: As $N \to \infty$, $t(N-K) \to N(0,1)$; this means that the standard normal distribution is the limiting, or asymptotic, distribution of the t-distribution.

The two-tail p-value for $z(\hat{\beta}_j)$ is labeled **P>|z|** in the output for the *Stata* **probit** and **dprobit** commands.