PPOL502-01, Spring 2016
Course Notes # 16: Binary Dependent Variable: Logit and Probit Models

---

Let's look at an application of these non-linear models in another example.

Example 2: Do individuals have accurate expectations about own HIV-positive status?

The data comes from a survey of nearly 600 women who turned up to receive antenatal care services at one clinic in Western Kenya in 2006.

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    positive |       456    .1973684    .3984499         0          1
hiv_expect~2 |       453    2.710817    .8813668         1          4
  agg_expect |       453    .3267108     .469529         0          1
         age |       456       24.75    6.450062        14         45
       agesq |       456    654.0746    360.3558       196       2025
-------------+--------------------------------------------------------
    doneprim |       453    .5916115    .4920791         0          1
     married |       456    .7763158    .4171705         0          1
     church1 |       456    3.348684    2.528762         0         12
       boils |       455    .7714286    .4203747         0          1
initial_li~k |       454    2.200441    3.734795         0         51
-------------+--------------------------------------------------------
roof_perma~t |       456    .7192982    .4498356         0          1
         loc |       456    .7697368    .4214636         0          1
```

The variable hiv_expect~2 captures the women's reports to the interviewer of their chances of being HIV-positive **before testing.** The variable takes on four values: 1 = High Chance 2=Moderate 3=Small 4=No Chance.
We want to examine if indeed reports about priors predict HIV-positive status. We can do this in three ways: LPM, Probit or Logit.

Let's start with the LPM.

```
. xi: reg positive i.hiv_expectations2;
i.hiv_expecta~2   _Ihiv_expec_1-4     (naturally coded; _Ihiv_expec_4 omitted)

      Source |       SS       df       MS              Number of obs =     453
-------------+------------------------------           F(  3,   449) =    4.46
       Model |  2.08889792     3  .696299307           Prob > F      =  0.0042
    Residual |  70.0303074   449  .155969504           R-squared     =  0.0290
-------------+------------------------------           Adj R-squared =  0.0225
       Total |  72.1192053   452  .159555764           Root MSE      =  .39493

------------------------------------------------------------------------------
    positive |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihiv_expe~1 |   .2296967   .0701551     3.27   0.001     .0918236    .3675698
_Ihiv_expe~2 |    .140411   .0619023     2.27   0.024     .0187568    .2620651
_Ihiv_expe~3 |   .0628248   .0529986     1.19   0.236    -.0413313    .1669808
       _cons |    .109589    .046223     2.37   0.018     .0187487    .2004294
```

The individual coefficients on the indicators for moderate and high chance of being HIV-positive are significant. But do these priors predict HIV-status? The overall F-stat answers this question. How do you interpret the constant, β_0?

What if we estimate this model using the *probit* command. But instead of reporting the coefficients from the main regression, we plot the marginal effects (at the means of Xs) using *dprobit*.

```
. xi: dprobit positive i.hiv_expectations2
i.hiv_expecta~2   _Ihiv_expec_1-4     (naturally coded; _Ihiv_expec_4 omitted)
Iteration 0:   log likelihood = -225.84803
Iteration 1:   log likelihood = -219.51987
Iteration 2:   log likelihood = -219.48784
Iteration 3:   log likelihood = -219.48784

Probit regression, reporting marginal effects           Number of obs =    453
                                                        LR chi2(3)    =  12.72
                                                        Prob > chi2   = 0.0053
Log likelihood = -219.48784                             Pseudo R2     = 0.0282
------------------------------------------------------------------------------
positive |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
_Ihiv_~1*|   .2720133   .0971522     3.13   0.002    .12362   .081598  .462428
_Ihiv_~2*|   .1712899     .08144     2.30   0.022   .203091    .01167  .330909
_Ihiv_~3*|   .0773179   .0587979     1.30   0.192   .512141  -.037924   .19256
---------+--------------------------------------------------------------------
  obs. P |   .1986755
 pred. P |   .1921443  (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
```

The overall likelihood-ratio (LR) test suggests that priors do predict HIV-status.
How do we interpret the coefficient on `_Ihiv_~1*`?

What is the baseline category?

We can also estimate the same regression using the logit command.

```
. xi: logit positive i.hiv_expectations2, or;
i.hiv_expecta~2   _Ihiv_expec_1-4     (naturally coded; _Ihiv_expec_4 omitted)

Iteration 0:   log likelihood = -225.84803
Iteration 1:   log likelihood = -219.72185
Iteration 2:   log likelihood =  -219.4881
Iteration 3:   log likelihood = -219.48784

Logistic regression                               Number of obs   =        453
                                                  LR chi2(3)      =      12.72
                                                  Prob > chi2     =     0.0053
Log likelihood = -219.48784                       Pseudo R2       =     0.0282
------------------------------------------------------------------------------
    positive | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihiv_expe~1 |   4.172297   1.957163     3.05   0.002     1.663754    10.46312
_Ihiv_expe~2 |   2.708333   1.206207     2.24   0.025     1.131367    6.483369
_Ihiv_expe~3 |   1.692708    .699133     1.27   0.203     .7533791    3.803214
```

The odds-ratio coefficients reported using the `logit, or` command are consistent with the results from the LPM and probit regressions. But the interpretation is very different.

For women reporting a high chance of being HIV-positive, there is a more than four-fold increase in the odds of being HIV-positive, relative to women reporting 'No chance' of being HIV-positive.

If we wanted to produce marginal effects coefficients akin to the `dprobit` command, we could run the `mfx compute` command.

```
. mfx compute

Marginal effects after logit
      y  = Pr(positive) (predict)
         =  .19054742
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
_Ihiv_~1*|    .286723      .10761    2.66   0.008   .075816   .49763    .12362
_Ihiv_~2*|   .1811625      .09062    2.00   0.046   .003559  .358766   .203091
_Ihiv_~3*|   .0809302       .0628    1.29   0.197  -.042153  .204014   .512141
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

We can now interpret the shaded coefficient in terms of changes in the likelihood of being HIV-positive (as opposed to the odds ratio).

Women reporting a high chance of being HIV-positive are nearly 29 percentage points more likely to test positive than women reporting 'No chance' of being positive.

But for researchers to be convinced that we can use reported priors to target an intervention, they would like to be able to know if there are other observable characteristics that can predict HIV-positive status. Those other variables may be much easier to collect and would thus constitute the new targeting rule.

In order to do this we include a range of additional independent variables to assess the extent to which there are lower cost signals of HIV-positive status.

```
. xi: reg positive i.hiv_expectations2 $controls;
i.hiv_expecta~2   _Ihiv_expec_1-4     (naturally coded; _Ihiv_expec_4 omitted)

      Source |       SS       df       MS              Number of obs =     447
-------------+------------------------------           F( 12,   434) =    2.45
       Model |  4.52641263    12  .377201053           Prob > F      =  0.0042
    Residual |  66.7532294   434  .153809284           R-squared     =  0.0635
-------------+------------------------------           Adj R-squared =  0.0376
       Total |  71.2796421   446  .159819825           Root MSE      =  .39219

------------------------------------------------------------------------------
    positive |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihiv_expe~1 |      .2111   .0716457     2.95   0.003     .0702844    .3519156
_Ihiv_expe~2 |   .1076026   .0634291     1.70   0.091    -.0170637     .232269
_Ihiv_expe~3 |    .035878   .0543554     0.66   0.510    -.0709545    .1427104
```

```
         age |   .0767969   .0237012     3.24   0.001     .0302135    .1233802
       agesq |  -.0013069   .0004176    -3.13   0.002    -.0021276   -.0004862
    doneprim |  -.0266096   .0402099    -0.66   0.508      -.10564    .0524207
     married |  -.0452684   .0517589    -0.87   0.382    -.1469977    .0564608
     church1 |   .0079856   .0076686     1.04   0.298    -.0070866    .0230578
       boils |    .003679   .0456762     0.08   0.936    -.0860951    .0934531
initial_li~k |  -.0057981   .0050158    -1.16   0.248    -.0156563    .0040601
roof_perma~t |  -.0122372   .0437512    -0.28   0.780    -.0982278    .0737534
         loc |   .0125603   .0444962     0.28   0.778    -.0748946    .1000151
       _cons |     -.8811   .3105085    -2.84   0.005    -1.491387   -.2708126
------------------------------------------------------------------------------
```

How would you interpret the coefficients on age, and age squared?

Are the priors still informative? The individual t-stats are no longer as large as before. But what happens if we formally test the null that priors are not informative?

```
.         test _Ihiv_expec_1 _Ihiv_expec_2 _Ihiv_expec_3;

 ( 1)  _Ihiv_expec_1 = 0
 ( 2)  _Ihiv_expec_2 = 0
 ( 3)  _Ihiv_expec_3 = 0

       F(  3,   434) =    3.87
            Prob > F =    0.0095
```

Adding observable characteristics does not seem to blunt the information contained in the reported priors.
How would we answer this question using probit and logit commands?

```
. xi: dprobit positive i.hiv_expectations2 $controls;
i.hiv_expecta~2   _Ihiv_expec_1-4     (naturally coded; _Ihiv_expec_4 omitted)

Iteration 0:   log likelihood = -223.12425
Iteration 1:   log likelihood = -208.90531
Iteration 2:   log likelihood = -208.70881
Iteration 3:   log likelihood = -208.70851

Probit regression, reporting marginal effects           Number of obs =    447
                                                        LR chi2(12)   =  28.83
                                                        Prob > chi2   = 0.0042
Log likelihood = -208.70851                             Pseudo R2     = 0.0646
------------------------------------------------------------------------------
positive |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
_Ihiv_~1*|    .239171   .0996194     2.72   0.007    .12528    .04392  .434421
_Ihiv_~2*|   .1277091   .0799176     1.74   0.082   .203579  -.028927  .284345
_Ihiv_~3*|   .0446034   .0593592     0.75   0.454   .514541  -.071738  .160945
     age |   .0854899   .0265324     3.17   0.002   24.8255   .033487  .137492
   agesq |  -.0014737    .000476    -3.05   0.002   658.141  -.002407 -.000541
doneprim*|  -.0335641   .0406738    -0.83   0.405   .588367  -.113283  .046155
 married*|   -.050665   .0582604    -0.91   0.365   .780761  -.164853  .063523
 church1 |   .0074991   .0075049     1.00   0.318   3.37136   -.00721  .022208
   boils*|   .0071544   .0453508     0.16   0.876   .778523  -.081731   .09604
initia~k |  -.0060785   .0056644    -1.07   0.284   2.18568  -.017181  .005024
roof_p~t*|  -.0066834    .043563    -0.15   0.877   .718121  -.092065  .078699
     loc*|   .0109311   .0441898     0.24   0.807   .769575  -.075679  .097542
---------+--------------------------------------------------------------------
  obs. P |   .1991051
 pred. P |   .1831471  (at x-bar)
------------------------------------------------------------------------------
```

The marginal effects for age and age squared have increased a little, and as with the LPM, the effects of priors are a little weaker than before. To formally test the hypothesis that priors are uninformative, we can use the analogous formula for the F-stat in the logit/probit world.

LR-stat $= -2(LL_{restricted} - LL_{unrestricted}) \sim \lambda_q$ where q is the number of restrictions, where $LL_{unrestricted}$ is the log-likelihood for the unrestricted model and $LL_{restricted}$ is the log-likelihood for the restricted model.

In order to implement this test, we need to run both the restricted and un-restricted models and store the estimates to be used later for the LR test.

Restricted Model

```
. xi: dprobit positive $controls

Iteration 0:   log likelihood = -223.12425
Iteration 1:   log likelihood = -214.05119
Iteration 2:   log likelihood = -213.93815
Iteration 3:   log likelihood = -213.93805

Probit regression, reporting marginal effects           Number of obs =    447
                                                        LR chi2(9)    =  18.37
                                                        Prob > chi2   = 0.0311
Log likelihood = -213.93805                             Pseudo R2     = 0.0412

------------------------------------------------------------------------------
positive |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
     age |    .090645   .0264353     3.37   0.001   24.8255   .038833  .142457
   agesq |  -.0015351   .0004735    -3.19   0.001   658.141  -.002463 -.000607
doneprim*|  -.0220733   .0403569    -0.55   0.582   .588367  -.101171  .057025
 married*|  -.0543892    .058498    -0.97   0.332   .780761  -.169043  .060265
 church1 |   .0055567   .0074552     0.74   0.456   3.37136  -.009055  .020169
   boils*|   .0076474   .0453549     0.17   0.867   .778523  -.081247  .096541
initia~k |  -.0061592   .0059882    -1.03   0.305   2.18568  -.017896  .005578
roof_p~t*|  -.0112732    .043997    -0.26   0.796   .718121  -.097506  .074959
     loc*|   .0058212   .0448834     0.13   0.897   .769575  -.082149  .093791
---------+--------------------------------------------------------------------
  obs. P |   .1991051
 pred. P |   .1879898  (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
```

Using the stored estimates we can run the LR test as follows:

```
. lrtest A b, stats

Likelihood-ratio test                                 LR chi2(3)  =     10.46
(Assumption: b nested in A)                           Prob > chi2 =    0.0150

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           b |    447   -223.1242    -213.938     10     447.8761    488.9017
           A |    447   -223.1242   -208.7085     13      443.417    496.7503
-----------------------------------------------------------------------------
```

We could have calculated our LR-stat by using the formula

LR-stat $= -2(LL_{restricted} - LL_{unrestricted}) \sim \lambda_q$

$= -2(\mathbf{-213.93805- -208.70851})$

$= -2*-5.23$

$= 10.46$

We can explore the relationship between age and the likelihood of being HIV-positive by using the mfx compute command.

```
. mfx compute, at(age=25, agesq=625)

warning: no value assigned in at() for variables doneprim married church1 boils
initial_livestock roof_permanent loc;
means used for doneprim married church1 boils initial_livestock roof_permanent loc

Marginal effects after dprobit
      y  = Pr(positive) (predict)
         =  .26175401
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
     age |   .1094389      .03721    2.94   0.003   .036518   .18236        25
   agesq |  -.0018534      .00066   -2.80   0.005   -.00315 -.000557       625
doneprim*|  -.0265997      .04883   -0.54   0.586  -.122296  .069097   .588367
 married*|   -.064758      .06971   -0.93   0.353  -.201392  .071876   .780761
 church1 |   .0067088      .00899    0.75   0.456   -.01092  .024337   3.37136
   boils*|    .009251      .05501    0.17   0.866  -.098563  .117065   .778523
initia~k |  -.0074362      .00721   -1.03   0.303  -.021576  .006704   2.18568
roof_p~t*|  -.0135797      .05283   -0.26   0.797  -.117128  .089968   .718121
     loc*|   .0070383      .05434    0.13   0.897  -.099471  .113547   .769575
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1

. mfx compute, at(age=20, agesq=400)

warning: no value assigned in at() for variables doneprim married church1 boils
initial_livestock roof_permanent loc;
means used for doneprim married church1 boils initial_livestock roof_permanent loc

Marginal effects after dprobit
      y  = Pr(positive) (predict)
         =  .14965733
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
     age |   .0782753      .02151    3.64   0.000   .036109  .120442        20
   agesq |  -.0013256       .0004   -3.34   0.001  -.002103 -.000548       400
doneprim*|  -.0190835      .03494   -0.55   0.585   -.08757  .049403   .588367
 married*|  -.0473766      .05049   -0.94   0.348  -.146333   .05158   .780761
 church1 |   .0047984      .00651    0.74   0.461   -.00796  .017557   3.37136
   boils*|   .0065958      .03906    0.17   0.866  -.069969  .083161   .778523
initia~k |  -.0053187       .0052   -1.02   0.306  -.015508  .004871   2.18568
roof_p~t*|  -.0097485      .03806   -0.26   0.798  -.084338  .064841   .718121
     loc*|   .0050224      .03865    0.13   0.897  -.070738  .080783   .769575
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```