ECON4004 – Lab 3 solutions

Question 1

(i) To calculate how main participated in the training program one can issue the commands

```
. count if train==1
  185

. count if train==0
  260
```

185 out of 445 participated in the job training program. The longest time in the experiment was 24 months (obtained from the variable *mosinex*, which denotes the number of months prior to January 78 in the experiment).

(ii) A regression of train on the explanatory variables gives

```
. regress train unem74 unem75 age educ black hisp married

      Source |       SS           df       MS      Number of obs   =       445
-------------+----------------------------------   F(7, 437)       =      1.43
       Model |  2.41922955         7  .345604222   Prob > F        =    0.1915
    Residual |  105.670658       437  .241809286   R-squared       =    0.0224
-------------+----------------------------------   Adj R-squared   =    0.0067
       Total |  108.089888       444  .243445693   Root MSE        =    .49174

------------------------------------------------------------------------------
       train |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      unem74 |     .02088   .0772939     0.27   0.787    -.1310341     .172794
      unem75 |  -.0955711   .0719021    -1.33   0.184     -.236888    .0457459
         age |   .0032057   .0034027     0.94   0.347     -.003482    .0098933
        educ |   .0120131   .0133419     0.90   0.368    -.0142092    .0382354
       black |  -.0816663   .0877325    -0.93   0.352    -.2540963    .0907637
        hisp |  -.2000168   .1169708    -1.71   0.088    -.4299122    .0298785
     married |   .0372887   .0644037     0.58   0.563    -.0892909    .1638683
       _cons |   .3380222   .1894451     1.78   0.075    -.0343147    .7103591
------------------------------------------------------------------------------
```

The *F* statistic for joint significance of the explanatory variables is $F(7,437) = 1.43$ with *p*-value = .19. Therefore, they are jointly insignificant at even the 15% level. Note that, even though we have estimated a linear probability model, the null hypothesis we are testing is that all slope coefficients are zero, and so there is no heteroskedasticity under H_0. This means that the usual *F* statistic is asymptotically valid.

(iii) We first estimate the model $P(train = 1|\mathbf{x}) = \Phi(\beta_0 + \beta_1 unem74 + \beta_2 unem75 + \beta_3 age + \beta_4 educ + \beta_5 black + \beta_6 hisp + \beta_7 married)$ by probit maximum likelihood, and obtain

```
. probit train unem74 unem75 age educ black hisp married

Iteration 0:   log likelihood =     -302.1
Iteration 1:   log likelihood = -297.01499
Iteration 2:   log likelihood =  -297.0088
Iteration 3:   log likelihood =  -297.0088

Probit regression                               Number of obs     =        445
                                                LR chi2(7)        =      10.18
                                                Prob > chi2       =     0.1785
Log likelihood =  -297.0088                     Pseudo R2         =     0.0169

------------------------------------------------------------------------------
       train |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      unem74 |   .0530256   .1992686     0.27   0.790    -.3375337    .4435849
      unem75 |  -.2477249     .18505    -1.34   0.181    -.6104163    .1149665
         age |   .0083443   .0087982     0.95   0.343    -.0088999    .0255886
        educ |   .0314431   .0343238     0.92   0.360    -.0358304    .0987165
       black |  -.2069299   .2249003    -0.92   0.358    -.6477264    .2338666
        hisp |  -.5397772   .3085029    -1.75   0.080    -1.144432    .0648773
     married |   .0966251   .1655823     0.58   0.560    -.2279101    .4211604
       _cons |  -.4241079   .4870267    -0.87   0.384    -1.378663    .5304469
------------------------------------------------------------------------------
```

In the probit model, the way to test whether the all included regressors are jointly significant is to perform a likelihood ratio test. This test compares the value of the likelihood when all regressors are included and with that when no regressors are included. The test statistic follows the chi-square distribution (denoted by χ^2), with degrees of freedom equal to the number of regressors. In our case, and as shown in the Stata output above, the likelihood ratio test for joint significance is 10.18. In a χ^2_7 distribution this gives *p*-value = .18, which is very similar to that obtained for the LPM in part (ii).

(iv) Training eligibility was randomly assigned among the participants, so it is not surprising that *train* appears to be independent of other observed factors. However, there can be a difference between eligibility and actual participation, as men can always refuse to participate if chosen.

(v) The simple LPM results are as follows:

```
. regress unem78 train, r

Linear regression                               Number of obs     =        445
                                                F(1, 443)         =       6.50
                                                Prob > F          =     0.0111
                                                R-squared         =     0.0139
                                                Root MSE          =     .45941

------------------------------------------------------------------------------
             |               Robust
      unem78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |  -.1106029   .0433918    -2.55   0.011    -.1958823   -.0253236
       _cons |   .3538462   .0297212    11.91   0.000      .295434    .4122583
------------------------------------------------------------------------------
```

Participating in the job training program lowers the estimated probability of being unemployed in 1978 by .111, or 11.1 percentage points. This is a large effect: the probability of being unemployed without participation is .354, and the training program reduces it to .243. The differences is statistically significant at almost the 1% level against at two-sided alternative. (Note that this is another case where, because training was randomly assigned, we have confidence that OLS is consistently estimating a causal effect, even though the *R*-squared from the regression is very small. There is much about being unemployed that we are not explaining, but we can be pretty confident that this job training program was beneficial.)

(vi) The estimated probit model is as follows:

```
. probit unem78 train, r

Iteration 0:   log pseudolikelihood = -274.73494
Iteration 1:   log pseudolikelihood = -271.58459
Iteration 2:   log pseudolikelihood =  -271.5828
Iteration 3:   log pseudolikelihood =  -271.5828

Probit regression                               Number of obs     =        445
                                                Wald chi2(1)      =       6.23
                                                Prob > chi2       =     0.0126
Log pseudolikelihood =  -271.5828               Pseudo R2         =     0.0115

------------------------------------------------------------------------------
             |               Robust
      unem78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |  -.3209508    .128621    -2.50   0.013    -.5730433   -.0688582
       _cons |  -.3749572   .0798356    -4.70   0.000     -.531432   -.2184824
------------------------------------------------------------------------------
```

where standard errors are in parentheses. It does not make sense to compare the coefficient on *train* for the probit, −.321, with the LPM estimate. The probabilities have different functional forms. However, note that the probit and LPM *t* statistics are essentially the same (although the LPM standard errors should be made robust to heteroskedasticity).

(vii) There are only two fitted values (i.e., predicted probabilities of *unem78*=1) in each case. In the LPM, they are equal to .354 when *train* = 0 and .243 when *train* = 1. In the probit model they are equal to $\Phi(-.3749572) = .354$ when *train* = 0 and $\Phi(-.3749572-.3209508)=.243$ when *train* = 1, where Φ denotes the cumulative normal distribution. Hence, fitted values are identical in both models. This has to be the case, because any method simply delivers the cell frequencies as the estimated probabilities. The LPM estimates are easier to interpret because they do not involve the transformation by $\Phi(\cdot)$, but it does not matter which is used provided the probability differences are calculated.

(viii) The fitted values are no longer going to be identical because the model is not saturated. That is, the explanatory variables are not an exhaustive, mutually exclusive set of dummy variables. To obtain the fitted probabilities from the LPM we run the following command:

```
. regress unem78 train unem74 unem75 age educ black hisp married, r

Linear regression                               Number of obs     =        445
                                                F(8, 436)         =       3.93
                                                Prob > F          =     0.0002
                                                R-squared         =     0.0462
                                                Root MSE          =     .45545

------------------------------------------------------------------------------
             |               Robust
      unem78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |  -.1117028   .0438196    -2.55   0.011    -.1978267   -.0255789
      unem74 |   .0386926   .0698225     0.55   0.580     -.098538    .1759231
      unem75 |   .0159613   .0654068     0.24   0.807    -.1125906    .1445132
         age |   .0000433   .0032717     0.01   0.989    -.0063869    .0064735
        educ |   .0001442   .0116097     0.01   0.990    -.0226737    .0229622
       black |   .1888328    .065795     2.87   0.004     .0595179    .3181477
        hisp |  -.0377011    .081827    -0.46   0.645    -.1985255    .1231234
     married |  -.0254373   .0591917    -0.43   0.668    -.1417739    .0908993
       _cons |   .1631823   .1615939     1.01   0.313    -.1544176    .4807822
------------------------------------------------------------------------------
```

To calculate the predicted probabilities, we used the predict command with the option xb, which calculates the linear index (hence the xb name), which in turn is equal to the predicted probability in the LPM. We name this predicted probability p_lpm, calculated as follows:

```
. predict p_lpm, xb
```

We then run the corresponding probit model

```
. probit unem78 train unem74 unem75 age educ black hisp married, r

Iteration 0:   log pseudolikelihood = -274.73494
Iteration 1:   log pseudolikelihood =  -263.3816
Iteration 2:   log pseudolikelihood =  -263.3128
Iteration 3:   log pseudolikelihood = -263.31279

Probit regression                               Number of obs     =        445
                                                Wald chi2(8)      =      22.97
                                                Prob > chi2       =     0.0034
Log pseudolikelihood = -263.31279               Pseudo R2         =     0.0416

------------------------------------------------------------------------------
             |               Robust
      unem78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |  -.3365897   .1306059    -2.58   0.010    -.5925726   -.0806068
      unem74 |    .106094   .2083582     0.51   0.611    -.3022806    .5144686
      unem75 |   .0636124   .1916636     0.33   0.740    -.3120414    .4392662
         age |   .0006757   .0093354     0.07   0.942    -.0176213    .0189728
        educ |  -.0018916   .0351258    -0.05   0.957    -.0707369    .0669538
       black |   .6336688   .2764519     2.29   0.022     .0918331    1.175504
        hisp |  -.1649409    .368144    -0.45   0.654      -.88649    .5566081
     married |   -.077768   .1788681    -0.43   0.664    -.4283431    .2728071
       _cons |  -1.010331   .5220632    -1.94   0.053    -2.033556    .0128939
------------------------------------------------------------------------------
```

To calculate the predicted probabilities from the probit we use again predict but with the option p, which denotes probability. If instead we used the option xb it would calculate again the value of the estimated linear index. We name this predicted probability from the probit as p_probit, and calculate it as follows:

```
. predict p_probit, p
```

When we correlated the two predicted probabilities, we obtain

```
. corr p_lpm p_probit
(obs=445)

             |    p_lpm p_probit
-------------+------------------
       p_lpm |   1.0000
    p_probit |   0.9932   1.0000
```

Hence, we observe a still very high correlation of .9932. This is due to the fact that the explanatory variables other than *train* are insignificant (with the exception of *black*). Therefore, the predicted probabilities are still primarily determined by *train*, and hence they are highly correlated.

(ix) To obtain the average partial effect of *train* using the probit model, we obtain fitted probabilities for each man for *train* = 1 and *train* = 0. Of course, one of these is a counterfactual, because the man was either in job training or not. Importantly, we evaluate the other regressors at their actual outcomes. The APE is the average, over all observations, of the differences in the estimated probabilities.

We evaluate the APE, using the margins command, which calculates marginal effect. Since *train* is a binary variable we can write it as ib0.train. This notation tells Stata that train is a binary variable (hence the i), and that its base value is 0. Typically Stata understands on its own that a variable is binary, but it is better to indicate it explicitly. We do the same thing for all binary variables in the specification. On the other hand, *age* and *educ* can be treated as continuous variables, and thus can be written as c.age and c.educ. Hence we first run the command

```
. probit unem78 ib0.train ib0.unem74 ib0.unem75 c.age c.educ ib0.black ib0.hisp ib0.married, r

Iteration 0:   log pseudolikelihood = -274.73494
Iteration 1:   log pseudolikelihood =  -263.3816
Iteration 2:   log pseudolikelihood =  -263.3128
Iteration 3:   log pseudolikelihood = -263.31279

Probit regression                               Number of obs     =        445
                                                Wald chi2(8)      =      22.97
                                                Prob > chi2       =     0.0034
Log pseudolikelihood = -263.31279               Pseudo R2         =     0.0416

------------------------------------------------------------------------------
             |               Robust
      unem78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.train |  -.3365897   .1306059    -2.58   0.010    -.5925726   -.0806068
    1.unem74 |    .106094   .2083582     0.51   0.611    -.3022806    .5144686
    1.unem75 |   .0636124   .1916636     0.33   0.740    -.3120414    .4392662
         age |   .0006757   .0093354     0.07   0.942    -.0176213    .0189728
        educ |  -.0018916   .0351258    -0.05   0.957    -.0707369    .0669538
     1.black |   .6336688   .2764519     2.29   0.022     .0918331    1.175504
      1.hisp |  -.1649409    .368144    -0.45   0.654      -.88649    .5566081
   1.married |   -.077768   .1788681    -0.43   0.664    -.4283431    .2728071
       _cons |  -1.010331   .5220632    -1.94   0.053    -2.033556    .0128939
------------------------------------------------------------------------------
```

To estimate the APE of train, we run the margins command as follows:

```
. margins, dydx(ib0.train)

Average marginal effects                        Number of obs     =        445
Model VCE    : Robust

Expression   : Pr(unem78), predict()
dy/dx w.r.t. : 1.train

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.train |  -.1123307   .0426718    -2.63   0.008    -.1959659   -.0286955
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Note: we would have obtained the same results if we used the command margins, dydx(train), as Stata understands that there is one variable named train and it is binary. It is safer, however, to use the full naming specification of *train*, that is, use margins,dydx(ib0.train)

With the variables in part (ii) appearing in the probit, the estimated APE is about −.112. Interestingly, rounded to three decimal places, this is the same as the coefficient on *train* in the linear regression. In other words, the linear probability model and probit give virtually the same estimated APEs.

(ix) To obtain the average partial effects of all regressors, we can use the command margins,dydx(*), where * tell Stata to cacluate the APE(AME) for all regressors. Hence we get

```
. margins, dydx(*)

Average marginal effects                          Number of obs     =        445
Model VCE    : Robust

Expression   : Pr(unem78), predict()
dy/dx w.r.t. : 1.train 1.unem74 1.unem75 age educ 1.black 1.hisp 1.married
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.train	-.1123307	.0426718	-2.63	0.008	-.1959659	-.0286955
1.unem74	.0353018	.0684546	0.52	0.606	-.0988667	.1694704
1.unem75	.0213189	.0640299	0.33	0.739	-.1041774	.1468153
age	.0002272	.0031391	0.07	0.942	-.0059253	.0063798
educ	-.000636	.011811	-0.05	0.957	-.0237851	.022513
1.black	.188783	.068846	2.74	0.006	.0538472	.3237188
1.hisp	-.0536882	.1155716	-0.46	0.642	-.2802044	.172828
1.married	-.0258306	.0586199	-0.44	0.659	-.1407235	.0890623

```
Note: dy/dx for factor levels is the discrete change from the base level.
```

We note that other than *train*, only being black has a statistically significant APE(AME), at increases on average the probability of being unemployed in 1978 by about 18.8 percentage points. We expect this result, as the coefficient of black was statistically significant in the probit regression. Almost always (i.e., with very few exceptions) a statistically significant probit coefficient will imply a statistically significant APE, and vice versa.

The result for *black* is very similar to the APE from the OLS regression, which is equal to the estimated coefficient. The remaining variables have statistically insignificant APES, with broadly similar patterns as the estimated OLS coefficients.

Question 2

(i) We first run the following probit

```
. probit vhappy ib0.occattend ib0.regattend ib1994.year, r

Iteration 0:   log pseudolikelihood = -10397.033
Iteration 1:   log pseudolikelihood =  -10339.48
Iteration 2:   log pseudolikelihood = -10339.463
Iteration 3:   log pseudolikelihood = -10339.463

Probit regression                               Number of obs     =     16,864
                                                Wald chi2(8)      =     115.69
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -10339.463               Pseudo R2         =     0.0055

------------------------------------------------------------------------------
             |               Robust
      vhappy |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.occattend |   .0122544   .0233063     0.53   0.599    -.0334251    .0579338
 1.regattend |   .3053249   .0300724    10.15   0.000     .2463841    .3642656
             |
        year |
       1996  |   .0482759   .0350063     1.38   0.168    -.0203352     .116887
       1998  |   .0798343   .0350279     2.28   0.023     .0111808    .1484878
       2000  |   .0894637   .0352215     2.54   0.011     .0204308    .1584966
       2002  |   .0455899   .0434216     1.05   0.294     -.039515    .1306947
       2004  |    .072181   .0435383     1.66   0.097    -.0131526    .1575146
       2006  |   .0638691   .0344714     1.85   0.064    -.0036936    .1314318
             |
       _cons |  -.6070756   .0262339   -23.14   0.000     -.658493   -.5556582
------------------------------------------------------------------------------
```

(note how we denote using 1994 as the base year).

We then calculate the APEs (average marginal effects) as follows:

```
. margins,dydx(*)

Average marginal effects                        Number of obs     =     16,864
Model VCE    : Robust

Expression   : Pr(vhappy), predict()
dy/dx w.r.t. : 1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.occattend |   .0042834    .008156     0.53   0.599    -.0117021    .0202688
 1.regattend |   .1122627   .0114661     9.79   0.000     .0897896    .1347358
             |
        year |
       1996  |    .016581   .0120244     1.38   0.168    -.0069864    .0401483
       1998  |   .0276457   .0121309     2.28   0.023     .0038696    .0514217
       2000  |   .0310558   .0122297     2.54   0.011      .007086    .0550256
       2002  |   .0156473   .0149673     1.05   0.296    -.0136881    .0449826
       2004  |   .0249465   .0151473     1.65   0.100    -.0047417    .0546347
       2006  |   .0220265   .0118825     1.85   0.064    -.0012628    .0453157
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Rounded to four decimal places, the APE for *occattend* is about .0043 (t = .53) and that for *regattend* is about .1123 (t = 9.79).

For a linear model estimated by OLS, we obtain the following results:

```
. regress vhappy ib0.occattend ib0.regattend ib1994.year, r

Linear regression                               Number of obs     =     16,864
                                                F(8, 16855)       =      13.58
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0071
                                                Root MSE          =     .45965

------------------------------------------------------------------------------
             |               Robust
      vhappy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.occattend |   .0042648    .008024     0.53   0.595    -.0114632    .0199928
 1.regattend |   .1121737   .0113857     9.85   0.000     .0898565     .134491
             |
        year |
       1996  |   .0167487    .012032     1.39   0.164    -.0068353    .0403327
       1998  |   .0278593   .0121477     2.29   0.022     .0040486      .05167
       2000  |   .0312657   .0122258     2.56   0.011      .007302    .0552295
       2002  |   .0157476   .0149857     1.05   0.293     -.013626    .0451211
       2004  |   .0251635   .0151638     1.66   0.097    -.0045591    .0548861
       2006  |   .0221839    .011884     1.87   0.062      -.00111    .0454779
             |
       _cons |   .2713457   .0088906    30.52   0.000     .2539191    .2887723
------------------------------------------------------------------------------
```

Hence, the APEs from a LPM are .0043 and .1122 (robust t = 9.85). So, they are very similar to the ones obtained using the probit.

(ii) We first generate the highinc variable using the following command

```
. ta income

 total family |
       income |      Freq.     Percent        Cum.
--------------+-----------------------------------
     lt $1000 |        176        1.17        1.17
 $1000 to 2999 |        182        1.21        2.38
 $3000 to 3999 |        150        1.00        3.38
 $4000 to 4999 |        156        1.04        4.41
 $5000 to 5999 |        209        1.39        5.80
 $6000 to 6999 |        202        1.34        7.15
 $7000 to 7999 |        218        1.45        8.59
 $8000 to 9999 |        399        2.65       11.25
$10000 - 14999 |      1,251        8.32       19.56
$15000 - 19999 |      1,099        7.30       26.87
$20000 - 24999 |      1,278        8.49       35.36
$25000 or more |      9,725       64.64      100.00
--------------+-----------------------------------
        Total |     15,045      100.00

. ta income, nol

      total |
     family |
     income |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        176        1.17        1.17
          2 |        182        1.21        2.38
          3 |        150        1.00        3.38
          4 |        156        1.04        4.41
          5 |        209        1.39        5.80
          6 |        202        1.34        7.15
          7 |        218        1.45        8.59
          8 |        399        2.65       11.25
          9 |      1,251        8.32       19.56
         10 |      1,099        7.30       26.87
         11 |      1,278        8.49       35.36
         12 |      9,725       64.64      100.00
------------+-----------------------------------
      Total |     15,045      100.00

. qui gen highinc=(income==12) if income<.
```

Adding the extra regressors, we then perform the following probit estimation:

```
. probit vhappy ib0.occattend ib0.regattend ib1994.year ib0.highinc ib0.unem10 c.educ c.teens, r

Iteration 0:   log pseudolikelihood = -5975.8765
Iteration 1:   log pseudolikelihood = -5820.2106
Iteration 2:   log pseudolikelihood = -5819.8741
Iteration 3:   log pseudolikelihood = -5819.8741

Probit regression                               Number of obs     =      9,768
                                                Wald chi2(12)     =     300.46
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -5819.8741               Pseudo R2         =     0.0261

------------------------------------------------------------------------------
             |               Robust
      vhappy |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.occattend |  -.0199884   .0309522    -0.65   0.518    -.0806535    .0406767
 1.regattend |   .2674814   .0400858     6.67   0.000     .1889146    .3460481
             |
        year |
       1996  |   .0359605   .0462361     0.78   0.437    -.0546605    .1265815
       1998  |     .05327   .0459723     1.16   0.247    -.0368341    .1433741
       2000  |    .088205   .0469093     1.88   0.060    -.0037356    .1801456
       2002  |  -.0523669    .057301    -0.91   0.361    -.1646749    .0599411
       2004  |   .0199594     .05804     0.34   0.731    -.0937968    .1337157
       2006  |  -.0181192   .0459246    -0.39   0.693    -.1081299    .0718914
             |
   1.highinc |   .3066568   .0310986     9.86   0.000     .2457048    .3676089
    1.unem10 |  -.2682503   .0297608    -9.01   0.000    -.3265804   -.2099201
        educ |   .0114743     .00493     2.33   0.020     .0018117     .021137
       teens |  -.0506173   .0279823    -1.81   0.070    -.1054616    .0042271
       _cons |  -.8438437   .0717174   -11.77   0.000    -.9844073   -.7032802
------------------------------------------------------------------------------
```

Before proceeding, we note that the estimation sample is now much smaller, due to missing values in income and *unem10*.

We then calculate the APEs (average marginal effects):

```
. margins,dydx(*)

Average marginal effects                          Number of obs     =      9,768
Model VCE    : Robust

Expression   : Pr(vhappy), predict()
dy/dx w.r.t. : 1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1.highi

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.occattend |  -.0067564   .0104412    -0.65   0.518    -.0272208     .013708
 1.regattend |   .0949556   .0147451     6.44   0.000     .0660558    .1238554
             |
        year |
       1996  |   .0121567   .0156335     0.78   0.437    -.0184845    .0427979
       1998  |   .0180866   .0156108     1.16   0.247      -.01251    .0486832
       2000  |   .0302029   .0160774     1.88   0.060    -.0013082    .0617141
       2002  |  -.0172918   .0188189    -0.92   0.358    -.0541762    .0195925
       2004  |   .0067199   .0195793     0.34   0.731    -.0316549    .0450947
       2006  |  -.0060395   .0153063    -0.39   0.693    -.0360393    .0239604
             |
   1.highinc |   .1019708   .0100237    10.17   0.000     .0823247    .1216169
    1.unem10 |  -.0891086   .0096003    -9.28   0.000    -.1079248   -.0702925
        educ |   .0038862   .0016685     2.33   0.020     .0006161    .0071563
       teens |  -.0171432   .0094726    -1.81   0.070    -.0357092    .0014228
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

We observe that the APE for *regattend* is about .0950 (t = 6.44). So, the APE estimate and its t statistic are somewhat lower when including the additional regressors, but it is still pretty large and very statistically significant. A person who reports attending a religious service regularly has, on average, almost a .10 higher probability of being "very happy."

(iii) The signs of the APEs of *highinc*, *unem*10, *educ*, and *teens* seem reasonable. Being in the highest income group (which, unfortunately, was not indexed to inflation) leads to about a .10 higher probability of being very happy, on average. Being unemployed in the past 10 years lowers the probability of being very happy by slightly less, about .09. Both are very statistically significant. Education has a slight positive effect: each year of education increase the probability of being very happy by about .004. Finally, having teenagers reduces the probability of being very happy. Each teenager is estimated to reduce the probability by about .017, although it is only marginally statistically significant.

(iv) If we add *black* and *female* to the probit from part (ii), we obtain

```
. probit vhappy ib0.occattend ib0.regattend ib1994.year ib0.highinc ib0.unem10 c.educ c.teens ib0.black ib0.female, r

Iteration 0:   log pseudolikelihood = -5975.8765
Iteration 1:   log pseudolikelihood =  -5813.317
Iteration 2:   log pseudolikelihood = -5812.9143
Iteration 3:   log pseudolikelihood = -5812.9143

Probit regression                               Number of obs     =      9,768
                                                Wald chi2(14)     =     310.38
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -5812.9143               Pseudo R2         =     0.0273
```

vhappy	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.occattend	-.011234	.0310918	-0.36	0.718	-.0721728	.0497047
1.regattend	.2803592	.0403941	6.94	0.000	.2011882	.3595303
year						
1996	.0397522	.0462319	0.86	0.390	-.0508606	.130365
1998	.0588921	.0459845	1.28	0.200	-.0312358	.1490201
2000	.0920765	.0469144	1.96	0.050	.0001259	.1840271
2002	-.0466923	.0573371	-0.81	0.415	-.159071	.0656864
2004	.022665	.0580557	0.39	0.696	-.091122	.136452
2006	-.0122806	.0459256	-0.27	0.789	-.1022931	.0777318
1.highinc	.2933894	.0315375	9.30	0.000	.2315771	.3552017
1.unem10	-.2647602	.0297972	-8.89	0.000	-.3231617	-.2063588
educ	.0102917	.0049369	2.08	0.037	.0006155	.0199679
teens	-.0456549	.0280588	-1.63	0.104	-.1006492	.0093395
1.black	-.1585431	.0427019	-3.71	0.000	-.2422373	-.0748489
1.female	.0046444	.0274705	0.17	0.866	-.0491967	.0584855
_cons	-.8124185	.0741288	-10.96	0.000	-.9577082	-.6671288

The APEs are calculated as follows:

```
. margins,dydx(*)

Average marginal effects                        Number of obs     =      9,768
Model VCE    : Robust

Expression   : Pr(vhappy), predict()
dy/dx w.r.t. : 1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1
               1.female

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.occattend |   -.003796   .0104944    -0.36   0.718    -.0243647    .0167726
 1.regattend |   .0995761   .0148751     6.69   0.000     .0704215    .1287308
             |
        year |
       1996  |   .0134091   .0155983     0.86   0.390     -.017163    .0439811
       1998  |   .0199608   .0155881     1.28   0.200    -.0105913    .0505128
       2000  |   .0314606   .0160452     1.96   0.050     .0000126    .0629085
       2002  |   -.015392   .0188097    -0.82   0.413    -.0522584    .0214745
       2004  |   .0076119   .0195416     0.39   0.697    -.0306889    .0459127
       2006  |  -.0040866   .0152816    -0.27   0.789     -.034038    .0258649
             |
   1.highinc |   .0975514   .0101856     9.58   0.000      .077588    .1175149
    1.unem10 |  -.0878733   .0096084    -9.15   0.000    -.1067053   -.0690412
        educ |   .0034814    .001669     2.09   0.037     .0002101    .0067527
       teens |  -.0154439   .0094879    -1.63   0.104    -.0340399    .0031522
     1.black |  -.0520126   .0135295    -3.84   0.000      -.07853   -.0254953
    1.female |   .0015709   .0092902     0.17   0.866    -.0166376    .0197793
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

In the probit regression, *black* is statistically significant ($t = -3.71$) while *female* is not ($t = .17$). The APE for *black* is about −.052, so that, other things in the model fixed, black people are, on average, .052 less likely to be very happy.

Adding an interaction between *black* and *female* we obtain (note that the interaction term can be written as ib0.black#ib0.female):

```
. probit vhappy ib0.occattend ib0.regattend ib1994.year ib0.highinc ib0.unem10 c.educ c.teens ib0.black ib0.female ib0.black#ib0.female, r

Iteration 0:   log pseudolikelihood = -5975.8765
Iteration 1:   log pseudolikelihood = -5812.7988
Iteration 2:   log pseudolikelihood = -5812.3758
Iteration 3:   log pseudolikelihood = -5812.3758

Probit regression                               Number of obs     =      9,768
                                                Wald chi2(15)     =     311.41
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -5812.3758               Pseudo R2         =     0.0274

------------------------------------------------------------------------------
             |               Robust
      vhappy |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.occattend |  -.0112967   .0310944    -0.36   0.716    -.0722406    .0496472
 1.regattend |   .2804282    .040397     6.94   0.000     .2012515     .359605
             |
        year |
       1996  |   .0405292   .0462255     0.88   0.381    -.0500712    .1311296
       1998  |   .0593751    .045978     1.29   0.197    -.0307401    .1494903
       2000  |   .0930056   .0469089     1.98   0.047     .0010658    .1849454
       2002  |   -.046499   .0573297    -0.81   0.417    -.1588631    .0658651
       2004  |   .0236477   .0580533     0.41   0.684    -.0901346      .13743
       2006  |  -.0120646   .0459179    -0.26   0.793     -.102062    .0779328
             |
   1.highinc |   .2921699    .031546     9.26   0.000     .2303408     .353999
    1.unem10 |  -.2646808   .0297977    -8.88   0.000    -.3230832   -.2062784
        educ |   .0103567   .0049357     2.10   0.036     .0006829    .0200305
       teens |  -.0448346   .0280778    -1.60   0.110    -.0998661    .0101969
     1.black |  -.1042254   .0675687    -1.54   0.123    -.2366576    .0282069
    1.female |   .0145224   .0290475     0.50   0.617    -.0424096    .0714544
             |
black#female |
        1 1  |  -.0894085   .0861391    -1.04   0.299    -.2582381    .0794211
             |
       _cons |  -.8183322   .0742853   -11.02   0.000    -.9639287   -.6727357
------------------------------------------------------------------------------
```

Writing the interaction term as ib0.black#ib0.female is important if we want to calculate the APEs, as Stata needs to know all the terms in the specification in which any particular variable shows up, so as to put it equal to 0 and 1 (when binary), or differentiate with respect to it (when continuous) correctly. If instead we had created a newly variable denoting the interaction term, then Stata would not know that this new variable consists of the interaction of the *black* and *female* variables.

We note from the probit results that the interaction term has a statistically insignificant *t* statistic, and the same is true for the *black* and *female* binary variables. This is likely due to the collinearity between the variables and their interaction. When we test the three dummies jointly we get

```
. testparm ib0.black ib0.female ib0.black#ib0.female

 ( 1)  [vhappy]1.black = 0
 ( 2)  [vhappy]1.female = 0
 ( 3)  [vhappy]1.black#1.female = 0

           chi2(  3) =   14.88
         Prob > chi2 =    0.0019
```

Hence, the three dummy variables are jointly very significant. It appears that a model with just *black* fits these data best.