Econometrics: Multiple Regression and Applications

ECON4004: LAB 3

Duong Trinh

University of Glasgow

February 21, 2024

Intro

- Duong Trinh
  - PhD Student in Economics (Bayesian Microeconometrics)
  - Email: Duong.Trinh@glasgow.ac.uk

- ECON4004-LB01
  - Wednesday 10am -12 pm
  - 5 sessions (7-Feb, 14-Feb, 21-Feb, 28-Feb, 6-March)
  - ST ANDREWS:357
- ECON4004-LB02
  - Wednesday 12-2 pm
  - 5 sessions (7-Feb, 14-Feb, 21-Feb, 28-Feb, 6-March)
  - ST ANDREWS:357

Record Attendance

Plan for LAB 3

- Exercise 1: based on Wooldridge, Exercise C17.8
- Exercise 2: based on Wooldridge, Exercise C17.14

- We will focus on *"Regression with Binary Dependent Variables"*
  - Linear Probability Model (LPM)
  - Probit Model (PROBIT)

Exercise 1: based on Wooldridge, Exercise C17.8

Picture the Scenario

- **Objective:** Examine the effect of *participation in the job training program* on *unemployment probabilities* and *earnings* in 1978.
- **Dataset:** `JTRAIN2.dta`
  - data on a job training experiment for a group of men. Men could enter the program starting in January 1976 through about mid-1977. The program ended in December 1977.
- **Key variables:**
  - `train`: job training indicator.
  - `unem78`: denoting being unemployed in 1978. (*outcome variable*)
  - `unem75`, `unem74`: denoting being unemployed in 1975 and 1974, respectively. (*pretraining variable*)
  - several demographic variables: `age`, `educ`, `black`, `hisp`, and `married`.

Questions

(i) How many men in the sample participated in the job training program? What was the highest number of months a man actually participated in the program?

(ii) Run a linear regression of `train` on `unem75`, `unem74`, `age`, `educ`, `black`, `hisp`, and `married`. Are these variables jointly significant at the 5% level?

(iii) Estimate a probit version of the linear model in part (ii). Compute the likelihood ratio test for joint significance of all variables. What do you conclude?

(iv) Based on your answers to parts (ii) and (iii), does it appear that participation in job training can be treated as exogenous for explaining 1978 unemployment status? Explain.

Questions

Single explanatory variable

(v) Run a simple regression of `unem78` on `train`. What is the estimated effect of participating in the job training program on the probability of being unemployed in 1978? Is it statistically significant?

(vi) Run a probit of `unem78` on `train`. Does it make sense to compare the probit coefficient on train with the coefficient obtained from the linear model in part (v)?

(vii) Find the fitted probabilities from parts (v) and (vi). Explain why they are identical. Which approach would you use to measure the effect and statistical significance of the job training program?

Questions

Additional controls & Average partial affect (APE)

(viii) Add all the variables from part (ii) as additional controls to the models from parts (v) and (vi). Are the fitted probabilities now identical? What is the correlation between them?

(ix) Using the model from part (viii), estimate the *average partial effect* of train on the 1978 unemployment probability.
How does the estimate compare with the OLS estimate from part (viii)?

(x) Estimate the *average partial effects* of the remaining regressors in (ix) on the 1978 unemployment probability.
How does the estimate compare with the OLS estimate from part (viii)?

(i) How many men in the sample participated in the job training program? What was the highest number of months man actually participated in the program?

- ◇ 185 men in the sample participated in the job training program. (»stata)
- ◇ The highest number of months a man actually participated in the program is 24. (»stata)

(ii) Run a linear regression of `train` on `unem75`, `unem74`, `age`, `educ`, `black`, `hisp`, and `married`.

Linear Probability Model (LPM)

$$\begin{aligned} \mathit{train}_i = \delta_0 + \delta_1 \cdot \mathit{unem74}_i + \delta_2 \cdot \mathit{unem75}_i + \delta_3 \cdot \mathit{age}_i + \delta_4 \cdot \mathit{educ}_i + \\ + \delta_5 \cdot \mathit{black}_i + \delta_6 \cdot \mathit{hisp}_i + \delta_7 \cdot \mathit{married}_i + u_i \end{aligned}$$

OLS estimation result (»stata)

$$\begin{aligned} \underset{(se)}{\widehat{\mathit{train}}} = \underset{(.189)}{.338} + \underset{(.077)}{.021} \cdot \mathit{unem74} - \underset{(.072)}{.096} \cdot \mathit{unem75} + \underset{(.003)}{.003} \cdot \mathit{age} + \underset{(.013)}{.012} \cdot \mathit{educ} + \\ - \underset{(.088)}{.082} \cdot \mathit{black} - \underset{(.112)}{.200} \cdot \mathit{hisp} + \underset{(.064)}{.037} \cdot \mathit{married} \end{aligned}$$

(ii) Are these variables jointly significant at the 5% level?

- Null Hypothesis: $H_0 : \delta_1 = \delta_2 = \ldots = \delta_7 = 0$ (»stata)
- 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) Estimate a probit version of linear model in part (ii).

Probit Model (PROBIT)

$$\Pr(\mathit{train} = 1 \mid \mathbf{x}) = \Phi(\delta_0 + \delta_1 \cdot \mathit{unem75} + \delta_2 \cdot \mathit{unem74} + \delta_3 \cdot \mathit{age} + \delta_4 \cdot \mathit{educ} + \\ + \delta_5 \cdot \mathit{black} + \delta_6 \cdot \mathit{hisp} + \delta_7 \cdot \mathit{married})$$

[SN] Stata command for Probit regression

```
* probit depvar indepvars [, options]
```

- For LPM, standard errors should be made robust to heteroskedasticity in order to yield more reliable results.
  - STATA: add option `robust` or `vce(robust)` to return heteroskedasticity-robust standard errors.

- For PROBIT, there is no such advantage in using robust standard errors. (reference)
  - They do not solve the problems associated with heteroskedasticity for a nonlinear model estimated using maximum likelihood, and have a different interpretation.
  - STATA: option `robust` or `vce(robust)` is unnecessary.

[SN] Stata command for Probit regression

```
* probit depvar indepvars [, options]
```

```
. 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	Coefficient	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

(iii) Estimate a probit version of linear model in part (ii).

Probit Model (PROBIT)

$$\Pr(train = 1 \mid \mathbf{x}) = \Phi(\delta_0 + \delta_1 \cdot unem75 + \delta_2 \cdot unem74 + \delta_3 \cdot age + \delta_4 \cdot educ + \\ + \delta_5 \cdot black + \delta_6 \cdot hisp + \delta_7 \cdot married)$$

Maximum Likelihood estimation result (»stata)

$$\underset{(se)}{\widehat{\Pr(train = 1 \mid \mathbf{x})}} = \Phi(\underset{(.487)}{-.424} + \underset{(.199)}{.053} \cdot unem74 - \underset{(.185)}{.247} \cdot unem75 + \underset{(.009)}{.008} \cdot age + \\ + \underset{(.034)}{.031} \cdot educ - \underset{(.225)}{.207} \cdot black - \underset{(.308)}{.540} \cdot hisp + \underset{(.166)}{.097} \cdot married)$$

(iii) Compute the likelihood ratio test for joint significance of all variables. What do you conclude?

- Null Hypothesis: $H_0 : \delta_1 = \delta_2 = \ldots = \delta_7 = 0$ (»stata)
- Likelihood ratio test for joint significance of all variables
- Idea: 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.
- 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) Based on your answers to parts (ii) and (iii), does it appear that participation in job training can be treated as exogenous for explaining 1978 unemployment status? Explain.

(»review)

- *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 (non-compliance issue).

(v) Run a simple regression of `unem78` on `train`.

Linear Probability Model (LPM)

$$unem78_i = \beta_0 + \beta_1 \cdot train_i + u_i$$

OLS estimation result (»stata)

$$\underset{(rb.se)}{\widehat{unem78}} = \underset{(.030)}{.354} - \underset{(.043)}{.111} \cdot train$$

(v) Run a simple regression of `unem78` on `train`.

Linear Probability Model (LPM)

$$unem78_i = \beta_0 + \beta_1 \cdot train_i + u_i$$

$$\Pr(unem78 = 1 \mid train) = E[unem78 \mid train] = \beta_0 + \beta_1 \cdot train$$

$\longrightarrow$ that's why we call "probabilily of..."

OLS estimation result (»stata)

$$\underset{(rb.se)}{\widehat{unem78}} = \underset{(.030)}{.354} - \underset{(.043)}{.111} \cdot train$$

$$\widehat{\Pr(unem78 = 1 \mid train)} = .354 - .111 \cdot train$$

(v) What is the estimated effect of participating in the job training program on the probability of being unemployed in 1978? Is it statistically significant?

Estimated Linear Probability Model (LPM) (»stata)

$$\widehat{unem78} = .354 - .111 \cdot train$$

- 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 differences is statistically significant at almost the 1% level against at two-sided alternative.
- 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. (»review)

(vi) Run a probit of `unem78` on `train`.

Probit Model (PROBIT)

$$\Pr(unem78 = 1 \mid train) = \Phi(\beta_0 + \beta_1 \cdot train)$$

Maximum Likelihood estimation result (»stata)

$$\underset{(se)}{\widehat{\Pr(unem78 = 1 \mid train)}} = \Phi(\underset{(.080)}{-.375} - \underset{(.128)}{.321} \cdot train)$$

(vi) Run a probit of `unem78` on `train`.

Probit Model (PROBIT)

$$\Pr(unem78 = 1 \mid train) = \Phi(\beta_0 + \beta_1 \cdot train)$$

$\Phi(\cdot)$ is CDF of the standard normal distribution that helps restrict returned values to [0,1].

Maximum Likelihood estimation result (»stata)

$$\underset{(se)}{\widehat{\Pr(unem78 = 1 \mid train)}} = \Phi(\underset{(.080)}{-.375} - \underset{(.128)}{.321} \cdot train)$$

(vi) Does it make sense to compare the probit coefficient on train with the coefficient obtained from the linear model in part (v)?

- ◇ It does not make sense to compare the coefficient on train for the probit ($-.321$) with the LPM estimate ($-.111$). 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) Find the fitted probabilities from parts (v) and (vi). Explain why they are identical.

Estimated Linear Probability Model (LPM)

$$\widehat{unem78} = .354 - .111 \cdot train$$

$$\widehat{\Pr(unem78 = 1 \mid train)} = .354 - .111 \cdot train$$

$\Rightarrow$ Predicted probabilities of being unemployed in 1978 (»stata)

- when $train = 0$ is: $\widehat{unem78}(train = 0) = \mathbf{.354}$
- when $train = 1$ is: $\widehat{unem78}(train = 1) = .354 - .111 = \mathbf{.243}$

(vii) Find the fitted probabilities from parts (v) and (vi). Explain why they are identical.

Estimated Probit Model (PROBIT)

$$\underset{(se)}{\widehat{\Pr(unem78 = 1 \mid train)}} = \Phi(\underset{(.080)}{-.375} - \underset{(.128)}{.321} \cdot train)$$

$\Rightarrow$ Predicted probabilities of being unemployed in 1978 (»stata)

- $\diamond$ when $train = 0$ is: $\widehat{\Pr(unem78 = 1 \mid train = 0)} = \Phi(-.375) = \mathbf{.354}$
- $\diamond$ when $train = 1$ is: $\widehat{\Pr(unem78 = 1 \mid train = 1)} = \Phi(-.375 - .321) = \mathbf{.243}$

(vii) Find the fitted probabilities from parts (v) and (vi). Explain why they are identical.

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 (here, we have only a single binary regressor). 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) Add all the variables from part (ii) as additional control to the models from parts (v) and (vi).

Linear Probability Model (LPM)

$$unem78_i = \beta_0 + \beta_1 \cdot train_i + \beta_2 \cdot unem74_i + \beta_3 \cdot unem75_i + \beta_4 \cdot age_i + \beta_5 \cdot educ_i + \\ + \beta_6 \cdot black_i + \beta_7 \cdot hisp_i + \beta_8 \cdot married_i + u_i$$

OLS estimation result (»stata)

$$\widehat{unem78} = .163 - .112 \cdot train + .039 \cdot unem74 + .016 \cdot unem75 + .000 \cdot age + \\ + .000 \cdot educ + .189 \cdot black - .038 \cdot hisp - .025 \cdot married$$

$$\widehat{\Pr(unem78 = 1 \mid \mathbf{x})} = .163 - .112 \cdot train + .039 \cdot unem74 + .016 \cdot unem75 + .000 \cdot age + \\ + .000 \cdot educ + .189 \cdot black - .038 \cdot hisp - .025 \cdot married$$

(viii) Add all the variables from part (ii) as additional control to the models from parts (v) and (vi).

Probit Model (PROBIT)

$$\Pr(\textit{unem78} = 1 \mid \mathbf{x}) = \Phi(\beta_0 + \beta_1 \cdot \textit{train}\beta_2 \cdot \textit{unem74} + \beta_3 \cdot \textit{unem75} + \beta_4 \cdot \textit{age} + \beta_5 \cdot \textit{educ} + \beta_6 \cdot \textit{black} + \beta_7 \cdot \textit{hisp} + \beta_8 \cdot \textit{married})$$

Maximum Likelihood estimation result (»stata)

$$\widehat{\Pr(\textit{unem78} = 1 \mid \mathbf{x})} = \Phi(-1.010 - .337 \cdot \textit{train}.106 \cdot \textit{unem74} + .064 \cdot \textit{unem75} + +.001 \cdot \textit{age} - .002 \cdot \textit{educ} + .634 \cdot \textit{black} - .165 \cdot \textit{hisp} - .078 \cdot \textit{married})$$

[SN] STATA command for Predicted probabilities

Linear Probability Model

```
* regress yvar xvar wvar1 wvar2 wvark, robust
* predict newvar, xb
// add option 'xb' to calculate linear index
```

Probit Model

```
* probit yvar xvar wvar1 wvar2 wvark
* predict newvar, p
// add option 'p' to calculate predicted probabilities
```

(viii) Are the fitted probabilities now identical? What is the correlation between them?

Linear Probability Model

```
* regress yvar xvar wvar1 wvar2 wvark, robust
* predict newvar, xb // add 'xb' to calculate linear index
```

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

. predict p_lpm, xb // predicted probability from LPM
```

Probit Model

```
* probit yvar xvar wvar1 wvar2 wvark
* predict newvar, p // add 'p' to calculate predicted probabilities
```

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

. predict p_probit, p // predicted probability from PROBIT
```

(viii) Are the fitted probabilities now identical? What is the correlation between them?

```
. summarize p_lpm p_probit
```

Variable	Obs	Mean	Std. dev.	Min	Max
p_lpm	445	.3078652	.0993342	-.0092491	.4105947
p_probit	445	.3077102	.1008801	.0550662	.4303571

```
* corr var1 var2 // return correlation coefficient
```

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

	p_lpm	p_probit
p_lpm	1.0000	
p_probit	0.9932	1.0000

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. Lower extreme values of predicted probabilities from LMP are even negative, while all values from probit fall in $[0, 1]$. However, we observe a still very high correlation of .9932.

[SN] STATA command for Average Partial Effects

```
* probit yvar ib0.binary_varname c.continuous_varname
// use 'c.' to explicitly indicate continuous variables
// use 'ib0.' to indicate binary variables, with base value 0
```

```
* probit yvar ib0.binary_var c.continuous_var
* margins, dydx(varname_of_interest)
// calculate APE for varname_of_interest among regressors.
```

```
* probit yvar ib0.binary_varname c.continuous_varname
* margins, dydx(*)
// use (*) to calculate APE for all regressors.
```

[SN] Probit regression - explicitly indicates types of variables

```
* probit yvar ib0.binary_varname c.continuous_varname
// use 'c.' to explicitly indicate continuous variables
// use 'ib0.' to indicate binary variables, with base value 0
```

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

Iteration 0:  Log likelihood = -274.73494
Iteration 1:  Log likelihood =  -263.3816
Iteration 2:  Log likelihood =  -263.3128
Iteration 3:  Log likelihood = -263.31279

Probit regression                                       Number of obs =    445
                                                        LR chi2(8)    =  22.84
                                                        Prob > chi2   = 0.0036
Log likelihood = -263.31279                             Pseudo R2     = 0.0416
```

unem78	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.train	-.3365897	.1316429	-2.56	0.011	-.5946051	-.0785744
1.unem74	.106094	.2125598	0.50	0.618	-.3105155	.5227035
1.unem75	.0636124	.1970995	0.32	0.747	-.3226956	.4499204
age	.0006757	.0091211	0.07	0.941	-.0172014	.0185529
educ	-.0018916	.0367938	-0.05	0.959	-.0740061	.0702229
1.black	.6336688	.2742692	2.31	0.021	.096111	1.171227
1.hisp	-.1649409	.3790471	-0.44	0.663	-.9078596	.5779777
1.married	-.077768	.1771557	-0.44	0.661	-.4249869	.2694509
_cons	-1.010331	.5380256	-1.88	0.060	-2.064842	.0441798

(ix) Using the model from part (viii), estimate the average partial effect of train on the 1978 unemployment probability. Compare with the OLS estimate from part (viii).

As `train` is a binary variable (»review)

$$
\begin{aligned}
APE_{train} = \frac{1}{n}\sum_{i=1}^{N} \Phi(&\hat{\beta}_0 + train\hat{\beta}_{train} + \\
&+ \text{sum of other regressors multiplied by their coefficients}) \\
&- \Phi(\hat{\beta}_0 + \text{sum of other regressors multiplied by their coefficients})]
\end{aligned}
$$

```
* probit yvar ib0.binary_var c.continuous_var
* margins, dydx(varname_of_interest)
// calculate APE for varname_of_interest among regressors.
```

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

. margins, dydx(ib0.train) // average partial effects for train with base value 0

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

Expression: Pr(unem78), predict()
dy/dx wrt:  1.train
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.train	$\boxed{-.1123307}$	.0429271	-2.62	$\boxed{0.009}$	-.1964663	-.0281951

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

(ix) Using the model from part (viii), estimate the average partial effect of train on the 1978 unemployment probability. Compare with the OLS estimate from part (viii).

- 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.

(x) Estimate the average partial effects of the remaining regressors in (ix) on the 1978 unemployment probability. Compare with the OLS estimate from part (viii).

```
* probit yvar ib0.binary_varname c.continuous_varname
* margins, dydx(*)
// use (*) to calculate APE for all regressors.
```

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

. margins, dydx(*) // average partial effects for all regressors

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

Expression: Pr(unem78), predict()
dy/dx wrt:  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	.0429271	-2.62	0.009	-.1964663	-.0281951
1.unem74	.0353018	.0699011	0.51	0.614	-.1017018	.1723055
1.unem75	.0213189	.0657959	0.32	0.746	-.1076387	.1502766
age	.0002272	.0030667	0.07	0.941	-.0057834	.0062379
educ	-.000636	.0123712	-0.05	0.959	-.024883	.023611
1.black	.188783	.0684525	2.76	0.006	.0546186	.3229474
1.hisp	-.0536882	.1188582	-0.45	0.651	-.286646	.1792697
1.married	-.0258306	.0580771	-0.44	0.656	-.1396597	.0879985

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

(x) Estimate the average partial effects of the remaining regressors in (ix) on the 1978 unemployment probability. Compare with the OLS estimate from part (viii).

- ◇ 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.

Exercise 2: based on Wooldridge, Exercise C17.14

Picture the Scenario

- **Objective:** Determinants of Happiness!
- **Dataset:** `happiness.dta`
  - contains independently pooled cross sections for the even years from 1994 through 2006, obtained from the General Social Survey.
- **Key variables:**
  - `vhappy`: a measure of "happiness", $= 1$ if the person reports being "very happy" and $= 0$ otherwise.
  - `occattend`: $= 1$ if attend religious services between several times a year and 2-3 times per month and $= 0$ otherwise.
  - `regattend`: $= 1$ if attend religious services more often that 2-3 times per month.
  - a full set of year dummies.

Questions

(i) Estimate a probit probability model relating `vhappy` to `occattend` and `regattend`. Find the average partial effects for `occattend` and `regattend`. How do these compare with those from estimating a linear probability model?

(ii) Include `highinc`, `unem10`, `educ`, and `teens` to the probit estimation in part (i). Is the APE of regattend affected much? What about its statistical significance?

(iii) Discuss the APEs and statistical significance of the four new variables in part (ii). Do the estimates make sense?

(iv) Controlling for the factors in part (ii), do there appear to be differences in happiness by gender or race? Justify your answer.

(i) Estimate a probit probability model relating `vhappy` to `occattend` and `regattend`.

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

Iteration 0:  Log likelihood = -10397.033
Iteration 1:  Log likelihood =  -10339.48
Iteration 2:  Log likelihood = -10339.463
Iteration 3:  Log likelihood = -10339.463

Probit regression                                       Number of obs = 16,864
                                                        LR chi2(8)    = 115.14
                                                        Prob > chi2   = 0.0000
Log likelihood = -10339.463                             Pseudo R2     = 0.0055
```

vhappy	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	.0122544	.0232981	0.53	0.599	-.0334091	.0579178
1.regattend	.3053249	.0300845	10.15	0.000	.2463604	.3642893
year						
1996	.0482759	.034976	1.38	0.168	-.0202759	.1168276
1998	.0798343	.0350035	2.28	0.023	.0112287	.1484398
2000	.0894637	.0352042	2.54	0.011	.0204648	.1584626
2002	.0455899	.0433746	1.05	0.293	-.0394227	.1306025
2004	.072181	.0435354	1.66	0.097	-.0131467	.1575087
2006	.0638691	.034432	1.85	0.064	-.0036165	.1313546
_cons	-.6070756	.0261378	-23.23	0.000	-.6583048	-.5558465

(i) Find the average partial effects for `occattend` and `regattend`.

```
. quiet probit vhappy ib0.occattend ib0.regattend ib1994.year

. margins,dydx(*)

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

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

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	.0042834	.0081532	0.53	0.599	-.0116965	.0202632
1.regattend	.1122627	.0114712	9.79	0.000	.0897796	.1347458
year						
1996	.016581	.0120143	1.38	0.168	-.0069667	.0401286
1998	.0276457	.0121232	2.28	0.023	.0038847	.0514066
2000	.0310558	.0122247	2.54	0.011	.0070959	.0550158
2002	.0156473	.0149513	1.05	0.295	-.0136567	.0449513
2004	.0249465	.015147	1.65	0.100	-.0047411	.0546342
2006	.0220265	.0118694	1.86	0.063	-.0012371	.04529

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

(i) How do these compare with those from estimating a linear probability model?

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

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

vhappy	Coefficient	Robust 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

(ii) Include `highinc`, `unem10`, `educ`, and `teens` to the probit estimation in part (i).

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

. margins, dydx(*)

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

Expression: Pr(vhappy), predict()
dy/dx wrt:  1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1.highinc 1.unem10
            educ teens
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	-.0067564	.0104435	-0.65	0.518	-.0272253	.0137125
1.regattend	.0949556	.0147601	6.43	0.000	.0660263	.1238848
year						
1996	.0121567	.0155867	0.78	0.435	-.0183927	.0427061
1998	.0180866	.0156145	1.16	0.247	-.0125173	.0486905
2000	.0302029	.0160702	1.88	0.060	-.001294	.0616999
2002	-.0172918	.0188304	-0.92	0.358	-.0541988	.0196152
2004	.0067199	.0195423	0.34	0.731	-.0315823	.0450222
2006	-.0060395	.0152607	-0.40	0.692	-.0359499	.0238709
1.highinc	.1019708	.0099953	10.20	0.000	.0823803	.1215613
1.unem10	-.0891086	.0096034	-9.28	0.000	-.107931	-.0702863
educ	.0038862	.0016398	2.37	0.018	.0006723	.0071
teens	-.0171432	.0094141	-1.82	0.069	-.0355946	.0013081

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

(ii) Is the APE of regattend affected much? What about its statistical significance?

We observe that the APE for regattend is about $.0950 (t = 6.43)$. 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) Discuss the APEs and statistical significance of the four new variables in part (ii).

The signs of the APEs of `highinc`, `unem10`, `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) Controlling for the factors in part (ii), do there appear to be differences in happiness by gender or race?

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

. margins, dydx(*)

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

Expression: Pr(vhappy), predict()
dy/dx wrt:  1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1.highinc 1.unem10 educ
            teens 1.black 1.female
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	-.003796	.0104925	-0.36	0.718	-.0243609	.0167688
1.regattend	.0995761	.0148764	6.69	0.000	.070419	.1287333
year						
1996	.0134091	.0155668	0.86	0.389	-.0171012	.0439194
1998	.0199608	.0156103	1.28	0.201	-.0106348	.0505563
2000	.0314606	.0160523	1.96	0.050	-1.36e-06	.0629225
2002	-.015392	.0188298	-0.82	0.414	-.0522977	.0215138
2004	.0076119	.0195077	0.39	0.696	-.0306224	.0458463
2006	-.0040866	.0152576	-0.27	0.789	-.033991	.0258178
1.highinc	.0975514	.0101496	9.61	0.000	.0776586	.1174443
1.unem10	-.0878733	.0096136	-9.14	0.000	-.1067156	-.0690309
educ	.0034814	.0016418	2.12	0.034	.0002636	.0066992
teens	-.0154439	.009423	-1.64	0.101	-.0339126	.0030248
1.black	-.0520126	.0135505	-3.84	0.000	-.0785711	-.0254542
1.female	.0015709	.0092531	0.17	0.865	-.0165649	.0197067

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

In the probit regression, black is statistically significant while female is not. 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.

(iv) Adding an interaction between `black` and `female`

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

. // include interaction term
. margins, dydx(*)

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

Expression: Pr(vhappy), predict()
dy/dx wrt:  1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1.highinc 1.unem10 educ teens 1.black 1.female
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	-.0038168	.0104917	-0.36	0.716	-.0243801	.0167465
1.regattend	.0995918	.0148764	6.69	0.000	.0704347	.1287489
year						
1996	.0136693	.0155676	0.88	0.380	-.0168427	.0441812
1998	.0201202	.0156091	1.29	0.197	-.0104732	.0507135
2000	.0317747	.0160547	1.98	0.048	.0003081	.0632413
2002	-.0153237	.0188266	-0.81	0.416	-.0522233	.0215759
2004	.0079413	.0195113	0.41	0.684	-.0303003	.0461828
2006	-.0040135	.0152543	-0.26	0.792	-.0339115	.0258844
1.highinc	.0971444	.0101577	9.56	0.000	.0772358	.1170531
1.unem10	-.0878395	.0096136	-9.14	0.000	-.1066819	-.0689971
educ	.0035031	.0016418	2.13	0.033	.0002853	.0067209
teens	-.015165	.0094268	-1.61	0.108	-.0336412	.0033112
1.black	-.0500396	.0137389	-3.64	0.000	-.0769674	-.0231118
1.female	.001447	.0092636	0.16	0.876	-.0167094	.0196034

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

(iv) Adding an interaction between black and female

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

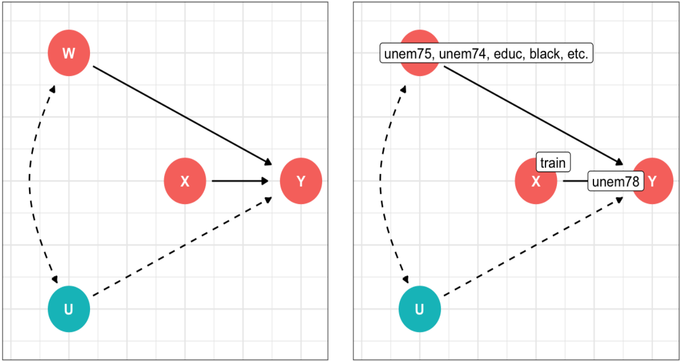
```
. quiet probit vhappy ib0.occattend ib0.regattend ib1994.year ib0.highinc ib0.unem10 c.educ c.teens ib0.black ib0.female ib0.black#ib0.female

. 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) =   14.78
         Prob > chi2 =    0.0020
```

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

BRIEF REVIEW

Causal Graph

Random Assignment



(»back1iv) (»back1v)

Average Partial Affect (APE)

For a continuous variable *cvar*

$$APE_{cvar} = \frac{1}{n}\sum_{i=1}^{N} \phi[(\hat{\beta}_0 + cvar\hat{\beta}_{cvar} +$$

$$+ \text{sum of other regressors multiplied by their coefficients}) \cdot \hat{\beta}_{cvar}]$$

For a binary variable *bvar*

$$APE_{bvar} = \frac{1}{n}\sum_{i=1}^{N} \Phi(\hat{\beta}_0 + bvar\hat{\beta}_{bvar} +$$

$$+ \text{sum of other regressors multiplied by their coefficients})$$

$$- \Phi(\hat{\beta}_0 + \text{sum of other regressors multiplied by their coefficients})]$$

STATA CODES & RESULTS

Exercise 1(i-I)

```
. tabulate train

     =1 if |
assigned to |
        job |
   training |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        260       58.43       58.43
          1 |        185       41.57      100.00
------------+-----------------------------------
      Total |        445      100.00

. count if train==1
  185
```

(»back1(i))

Exercise 1(i-II)

```
. summarize mosinex
```

Variable	Obs	Mean	Std. dev.	Min	Max
mosinex	445	18.1236	5.311937	5	24

```
. tabstat mosinex, s(max)
```

Variable	Max
mosinex	24

(»back1(i))

Exercise 1(ii)

```
. 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 | Coefficient  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
------------------------------------------------------------------------------
```

Exercise 1(ii)

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

Linear regression                               Number of obs     =        445
                                                F(7, 437)         =       1.60
                                                Prob > F          =     0.1334
                                                R-squared         =     0.0224
                                                Root MSE          =     .49174
```

train	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
unem74	.02088	.0772497	0.27	0.787	-.1309472	.1727072
unem75	-.0955711	.0722763	-1.32	0.187	-.2376234	.0464813
age	.0032057	.0033869	0.95	0.344	-.003451	.0098624
educ	.0120131	.0138597	0.87	0.387	-.0152268	.039253
black	-.0816663	.0888047	-0.92	0.358	-.2562038	.0928712
hisp	-.2000168	.1132098	-1.77	0.078	-.4225202	.0224865
married	.0372887	.0650005	0.57	0.566	-.0904638	.1650412
_cons	.3380222	.1944555	1.74	0.083	-.044162	.7202064

Exercise 1(iii)

```
. 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	Coefficient	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

Exercise 1(v)

```
. regress unem78 train, robust

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

unem78	Coefficient	Robust 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

Exercise 1(vi)

```
. probit unem78 train

Iteration 0:  Log likelihood = -274.73494
Iteration 1:  Log likelihood = -271.58459
Iteration 2:  Log likelihood =  -271.5828
Iteration 3:  Log likelihood =  -271.5828

Probit regression                                       Number of obs =    445
                                                        LR chi2(1)    =   6.30
                                                        Prob > chi2   = 0.0120
Log likelihood = -271.5828                              Pseudo R2     = 0.0115
```

unem78	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
train	-.3209508	.1284764	-2.50	0.012	-.5727599	-.0691416
_cons	-.3749572	.0797458	-4.70	0.000	-.5312561	-.2186583

Exercise 1(vii-I)

```
. regress unem78 train, robust coeflegend

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

------------------------------------------------------------------------------
      unem78 | Coefficient  Legend
-------------+----------------------------------------------------------------
       train |  -.1106029  _b[train]
       _cons |   .3538462  _b[_cons]
------------------------------------------------------------------------------

. display "From LPM, probability when train=0 is: " _b[_cons]
From LPM, probability when train=0 is: .35384615

. display "From LPM, probability when train=1 is: " _b[_cons]+_b[train]
From LPM, probability when train=1 is: .24324324
```

Exercise 1(vii-II)

```
. probit unem78 train, coeflegend

Iteration 0:  Log likelihood = -274.73494
Iteration 1:  Log likelihood = -271.58459
Iteration 2:  Log likelihood =  -271.5828
Iteration 3:  Log likelihood =  -271.5828

Probit regression                                       Number of obs =    445
                                                        LR chi2(1)    =   6.30
                                                        Prob > chi2   = 0.0120
Log likelihood = -271.5828                              Pseudo R2     = 0.0115

------------------------------------------------------------------------------
      unem78 | Coefficient  Legend
-------------+----------------------------------------------------------------
       train |  -.3209508  _b[train]
       _cons |  -.3749572  _b[_cons]
------------------------------------------------------------------------------

. display "From probit, probability when train=1 is: " normal(_b[_cons])
From probit, probability when train=1 is: .35384615

. display "From probit, probability when train=0 is: " normal(_b[_cons]+_b[train])
From probit, probability when train=0 is: .24324324
```

Exercise 1(viii-I)

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

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

unem78	Coefficient	Robust 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

Exercise 1(viii-II)

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

Iteration 0:   Log likelihood = -274.73494
Iteration 1:   Log likelihood =  -263.3816
Iteration 2:   Log likelihood =  -263.3128
Iteration 3:   Log likelihood = -263.31279

Probit regression                                       Number of obs =    445
                                                        LR chi2(8)    =  22.84
                                                        Prob > chi2   = 0.0036
Log likelihood = -263.31279                             Pseudo R2     = 0.0416
```

unem78	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
train	-.3365897	.1316429	-2.56	0.011	-.5946051	-.0785744
unem74	.106094	.2125598	0.50	0.618	-.3105155	.5227035
unem75	.0636124	.1970995	0.32	0.747	-.3226956	.4499204
age	.0006757	.0091211	0.07	0.941	-.0172014	.0185529
educ	-.0018916	.0367938	-0.05	0.959	-.0740061	.0702229
black	.6336688	.2742692	2.31	0.021	.096111	1.171227
hisp	-.1649409	.3790471	-0.44	0.663	-.9078596	.5779777
married	-.077768	.1771557	-0.44	0.661	-.4249869	.2694509
_cons	-1.010331	.5380256	-1.88	0.060	-2.064842	.0441798

Exercise 1(viii-III)

```
* regress yvar xvar wvar1 wvar2 wvark, robust
* predict newvar, xb // add 'xb' to calculate linear index
```

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

. predict p_lpm, xb // predicted probability from LPM
```

```
* probit yvar xvar wvar1 wvar2 wvark
* predict newvar, p // add 'p' to calculate predicted probabilities
```

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

. predict p_probit, p // predicted probability from PROBIT
```

```
* corr var1 var2 // return correlation coefficient
```

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

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

Exercise 1(ix)

```
* probit yvar ib0.binary_var c.continuous_var
* margins, dydx(varname_of_interest)
// calculate APE for varname_of_interest among regressors.
```

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

. margins, dydx(ib0.train) // average partial effects for train with base value 0

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

Expression: Pr(unem78), predict()
dy/dx wrt:  1.train
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.train	-.1123307	.0429271	-2.62	0.009	-.1964663	-.0281951

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

Exercise 1(x)

```
* probit yvar ib0.binary_varname c.continuous_varname
* margins, dydx(*)
// use (*) to calculate APE for all regressors.
```

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

. margins, dydx(*) // average partial effects for all regressors

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

Expression: Pr(unem78), predict()
dy/dx wrt:  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	.0429271	-2.62	0.009	-.1964663	-.0281951
1.unem74	.0353018	.0699011	0.51	0.614	-.1017018	.1723055
1.unem75	.0213189	.0657959	0.32	0.746	-.1076387	.1502766
age	.0002272	.0030667	0.07	0.941	-.0057834	.0062379
educ	-.000636	.0123712	-0.05	0.959	-.024883	.023611
1.black	.188783	.0684525	2.76	0.006	.0546186	.3229474
1.hisp	-.0536882	.1188582	-0.45	0.651	-.286646	.1792697
1.married	-.0258306	.0580771	-0.44	0.656	-.1396597	.0879985

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

Exercise 2(i-I)

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

Linear regression

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

vhappy	Coefficient	Robust 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

Exercise 2(i-II)

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

Iteration 0:  Log likelihood = -10397.033
Iteration 1:  Log likelihood =  -10339.48
Iteration 2:  Log likelihood = -10339.463
Iteration 3:  Log likelihood = -10339.463

Probit regression                                       Number of obs = 16,864
                                                        LR chi2(8)    = 115.14
                                                        Prob > chi2   = 0.0000
Log likelihood = -10339.463                             Pseudo R2     = 0.0055
```

vhappy	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	.0122544	.0232981	0.53	0.599	-.0334091	.0579178
1.regattend	.3053249	.0300845	10.15	0.000	.2463604	.3642893
year						
1996	.0482759	.034976	1.38	0.168	-.0202759	.1168276
1998	.0798343	.0350035	2.28	0.023	.0112287	.1484398
2000	.0894637	.0352042	2.54	0.011	.0204648	.1584626
2002	.0455899	.0433746	1.05	0.293	-.0394227	.1306025
2004	.072181	.0435354	1.66	0.097	-.0131467	.1575087
2006	.0638691	.034432	1.85	0.064	-.0036165	.1313546
_cons	-.6070756	.0261378	-23.23	0.000	-.6583048	-.5558465

Exercise 2(i-III)

```
. quiet probit vhappy ib0.occattend ib0.regattend ib1994.year

. margins,dydx(*)

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

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

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	.0042834	.0081532	0.53	0.599	-.0116965	.0202632
1.regattend	.1122627	.0114712	9.79	0.000	.0897796	.1347458
year						
1996	.016581	.0120143	1.38	0.168	-.0069667	.0401286
1998	.0276457	.0121232	2.28	0.023	.0038847	.0514066
2000	.0310558	.0122247	2.54	0.011	.0070959	.0550158
2002	.0156473	.0149513	1.05	0.295	-.0136567	.0449513
2004	.0249465	.015147	1.65	0.100	-.0047411	.0546342
2006	.0220265	.0118694	1.86	0.063	-.0012371	.04529

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

Exercise 2(ii-I)

```
. tab income, miss // table of frequencies, treat missing values like other values
```

total family income	Freq.	Percent	Cum.
lt $1000	176	1.03	1.03
$1000 to 2999	182	1.06	2.09
$3000 to 3999	150	0.88	2.96
$4000 to 4999	156	0.91	3.87
$5000 to 5999	209	1.22	5.09
$6000 to 6999	202	1.18	6.27
$7000 to 7999	218	1.27	7.55
$8000 to 9999	399	2.33	9.87
$10000 - 14999	1,251	7.30	17.17
$15000 - 19999	1,099	6.41	23.59
$20000 - 24999	1,278	7.46	31.04
$25000 or more	9,725	56.75	87.79
.	2,092	12.21	100.00
Total	17,137	100.00	

```
. tab income, miss nolabel //  display numeric codes rather than value labels
```

total family income	Freq.	Percent	Cum.
1	176	1.03	1.03
2	182	1.06	2.09
3	150	0.88	2.96
4	156	0.91	3.87
5	209	1.22	5.09
6	202	1.18	6.27
7	218	1.27	7.55
8	399	2.33	9.87
9	1,251	7.30	17.17
10	1,099	6.41	23.59
11	1,278	7.46	31.04
12	9,725	56.75	87.79
.	2,092	12.21	100.00
Total	17,137	100.00	

```
. gen highinc = (income==12) if (income != .) // only generate values for nonmissing
(2,092 missing values generated)

. tab highinc, miss
```

highinc	Freq.	Percent	Cum.
0	5,320	31.04	31.04
1	9,725	56.75	87.79
.	2,092	12.21	100.00
Total	17,137	100.00	

Exercise 2(ii-II)

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

. margins, dydx(*)

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

Expression: Pr(vhappy), predict()
dy/dx wrt:  1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1.highinc 1.unem10
            educ teens
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	-.0067564	.0104435	-0.65	0.518	-.0272253	.0137125
1.regattend	.0949556	.0147601	6.43	0.000	.0660263	.1238848
year						
1996	.0121567	.0155867	0.78	0.435	-.0183927	.0427061
1998	.0180866	.0156145	1.16	0.247	-.0125173	.0486905
2000	.0302029	.0160702	1.88	0.060	-.001294	.0616999
2002	-.0172918	.0188304	-0.92	0.358	-.0541988	.0196152
2004	.0067199	.0195423	0.34	0.731	-.0315823	.0450222
2006	-.0060395	.0152607	-0.40	0.692	-.0359499	.0238709
1.highinc	.1019708	.0099953	10.20	0.000	.0823803	.1215613
1.unem10	-.0891086	.0096034	-9.28	0.000	-.107931	-.0702863
educ	.0038862	.0016398	2.37	0.018	.0006723	.0071
teens	-.0171432	.0094141	-1.82	0.069	-.0355946	.0013081

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

Exercise 2(iv-I)

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

. margins, dydx(*)

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

Expression: Pr(vhappy), predict()
dy/dx wrt:  1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1.highinc 1.unem10 educ
            teens 1.black 1.female
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	-.003796	.0104925	-0.36	0.718	-.0243609	.0167688
1.regattend	.0995761	.0148764	6.69	0.000	.070419	.1287333
year						
1996	.0134091	.0155668	0.86	0.389	-.0171012	.0439194
1998	.0199608	.0156103	1.28	0.201	-.0106348	.0505563
2000	.0314606	.0160523	1.96	0.050	-1.36e-06	.0629225
2002	-.015392	.0188298	-0.82	0.414	-.0522977	.0215138
2004	.0076119	.0195077	0.39	0.696	-.0306224	.0458463
2006	-.0040866	.0152576	-0.27	0.789	-.033991	.0258178
1.highinc	.0975514	.0101496	9.61	0.000	.0776586	.1174443
1.unem10	-.0878733	.0096136	-9.14	0.000	-.1067156	-.0690309
educ	.0034814	.0016418	2.12	0.034	.0002636	.0066992
teens	-.0154439	.009423	-1.64	0.101	-.0339126	.0030248
1.black	-.0520126	.0135505	-3.84	0.000	-.0785711	-.0254542
1.female	.0015709	.0092531	0.17	0.865	-.0165649	.0197067

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

Exercise 2(iv-II)

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

. // include interaction term
. margins, dydx(*)

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

Expression: Pr(vhappy), predict()
dy/dx wrt:  1.occattend 1.regattend 1996.year 1998.year 2000.year 2002.year 2004.year 2006.year 1.highinc 1.unem10 educ teens 1.black 1.female
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1.occattend	-.0038168	.0104917	-0.36	0.716	-.0243801	.0167465
1.regattend	.0995918	.0148764	6.69	0.000	.0704347	.1287489
year						
1996	.0136693	.0155676	0.88	0.380	-.0168427	.0441812
1998	.0201202	.0156091	1.29	0.197	-.0104732	.0507135
2000	.0317747	.0160547	1.98	0.048	.0003081	.0632413
2002	-.0153237	.0188266	-0.81	0.416	-.0522233	.0215759
2004	.0079413	.0195113	0.41	0.684	-.0303003	.0461828
2006	-.0040135	.0152543	-0.26	0.792	-.0339115	.0258844
1.highinc	.0971444	.0101577	9.56	0.000	.0772358	.1170531
1.unem10	-.0878395	.0096136	-9.14	0.000	-.1066819	-.0689971
educ	.0035031	.0016418	2.13	0.033	.0002853	.0067209
teens	-.015165	.0094268	-1.61	0.108	-.0336412	.0033112
1.black	-.0500396	.0137389	-3.64	0.000	-.0769674	-.0231118
1.female	.001447	.0092636	0.16	0.876	-.0167094	.0196034

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

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.

Exercise 2(iv-III)

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

. 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) =   14.78
         Prob > chi2 =    0.0020
```