Exercise 8

The file **JTRAIN2.dta** contains 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. The idea is to test whether participation in the job training program had an effect on unemployment probabilities and earnings in 1978.

(i) The variable *train* is the job training indicator. 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?

smpl train –restrict
smpl full
summary mostrn

185 out of 445 participated in the job training program. The longest time in the experiment was 24 months.

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

ols train const unem74 unem75 age educ black hisp married

```
Model 1: OLS, using observations 1-445
Dependent variable: train

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        0.338022     0.189445      1.784    0.0751   *
  unem74       0.0208800    0.0772939     0.2701   0.7872
  unem75      -0.0955711    0.0719021    -1.329    0.1845
  age          0.00320567   0.00340269    0.9421   0.3467
  educ         0.0120131    0.0133419     0.9004   0.3684
  black       -0.0816663    0.0877325    -0.9309   0.3524
  hisp        -0.200017     0.116971     -1.710    0.0880   *
  married      0.0372887    0.0644037     0.5790   0.5629

Mean dependent var   0.415730   S.D. dependent var   0.493402
Sum squared resid    105.6707   S.E. of regression   0.491741
R-squared            0.022382   Adjusted R-squared   0.006722
F(7, 437)            1.429243   P-value(F)           0.191523
Log-likelihood      -311.5289   Akaike criterion     639.0579
Schwarz criterion    671.8425   Hannan-Quinn         651.9854
```

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

probit train const unem74 unem75 age educ black hisp married

```
Model 2: Probit, using observations 1-445
Dependent variable: train
Standard errors based on Hessian

             coefficient   std. error      z         slope
  ---------------------------------------------------------
  const      -0.424108     0.487027     -0.8708
  unem74      0.0530256    0.199269      0.2661    0.0206009
  unem75     -0.247725     0.185050     -1.339    -0.0969931
  age         0.00834435   0.00879823    0.9484    0.00325117
  educ        0.0314431    0.0343238     0.9161    0.0122510
  black      -0.206930     0.224900     -0.9201   -0.0815074
  hisp       -0.539777     0.308503     -1.750    -0.193576
  married     0.0966251    0.165582      0.5835    0.0378769

Mean dependent var   0.415730   S.D. dependent var   0.493402
McFadden R-squared   0.016853   Adjusted R-squared  -0.009629
Log-likelihood      -297.0088   Akaike criterion     610.0176
Schwarz criterion    642.8022   Hannan-Quinn         622.9452

Number of cases 'correctly predicted' = 266 (59.8%)
f(beta'x) at mean of independent vars = 0.390
Likelihood ratio test: Chi-square(7) = 10.1824 [0.1785]
```

After estimating the model P(train=1|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, the likelihood ratio test for joint significance is 10.18. In a χ^2 distribution this gives P-value =0.18, which is very similar to that obtained in 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.

Training eligibility was randomly assigned among the participants, so it is not surprising that *train* appears to be independent of other observed factors. Therefore, running a regression of train on unem78 would not suffer with the endogeneity issue (However, there can be a difference between eligibility and actual participation, as men can always refuse to participate if chosen.)

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

ols unem78 const train

```
Model 3: OLS, using observations 1-445
Dependent variable: unem78

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       0.353846     0.0284917    12.42     1.35e-030 ***
  train      -0.110603     0.0441888    -2.503    0.0127     **

Mean dependent var   0.307865   S.D. dependent var   0.462130
Sum squared resid    93.50021   S.E. of regression   0.459414
R-squared            0.013945   Adjusted R-squared   0.011719
F(1, 443)            6.264831   P-value(F)           0.012675
Log-likelihood      -284.3030   Akaike criterion     572.6061
Schwarz criterion    580.8022   Hannan-Quinn         575.8380
```

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) 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)?

```
? probit unem78 const train

Model 4: Probit, using observations 1-445
Dependent variable: unem78
Standard errors based on Hessian

             coefficient   std. error     z        slope
  -----------------------------------------------------------
  const       -0.374957    0.0797458    -4.702
  train       -0.320951    0.128476     -2.498   -0.110603

Mean dependent var   0.307865   S.D. dependent var   0.462130
McFadden R-squared   0.011473   Adjusted R-squared   0.004194
Log-likelihood      -271.5828   Akaike criterion     547.1656
Schwarz criterion    555.3618   Hannan-Quinn         550.3975

Number of cases 'correctly predicted' = 308 (69.2%)
f(beta'x) at mean of independent vars = 0.351
Likelihood ratio test: Chi-square(1) = 6.30427 [0.0120]
```

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) 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?

There are only two fitted values in each case, and they are the same: .354 when *train* = 0 and .243 when *train* = 1. 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.

$$\boldsymbol{P(Y=1|X)=\phi(-0.37-0.32)=\phi(-0.69)=0.245}$$

$$\boldsymbol{P(Y=0|X)=\phi(-0.37)=0.355}$$

(viii) Add all of 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?

ols unem78 const train unem74 unem75 age educ black hisp married
series yhat=$yhat
probit unem78 const train unem74 unem75 age educ black hisp married

```
series yhat2=$yhat
corr yhat yhat2
```

The fitted values are no longer identical because the model is not saturated, that is, the explanatory variables are not an exhaustive, mutually exclusive set of dummy variables. But, because the other explanatory variables are insignificant, the fitted values are highly correlated: the LPM and probit fitted values have a correlation of about .993