

Problem set 4

Your answers are due on November 7, 2023, noon. You should submit them through Gradescope.

1. (a) Let *grad* be a dummy variable for whether a student-athlete at a large university graduates in five years. Let *hsGPA* and *SAT* be high school grade point average and SAT score, respectively. Let *study* be the number of hours spent per week in an organized study hall. Suppose that, using data on 420 student-athletes, the following probit model is obtained

$$\begin{aligned} P(\text{grad} = 1 | \text{hsGPA}, \text{SAT}, \text{study}) &= \\ &= \Phi(-0.73 + 0.15 \times \text{hsGPA} + 0.00036 \times \text{SAT} + 0.046 \times \text{study}) \end{aligned}$$

Holding *hsGPA* fixed at 3.0 and *SAT* fixed at 1200, compute the estimated difference in graduation probability for someone who spent 10 hours per week in study hall and someone who spent 5 hours per week.

- (b) Assume one estimates the following probit model:

$$P(\text{employed} = 1 | \text{age}, \text{educ}) = \Phi(\alpha + \beta \text{age} + \gamma \text{age}^2 + \delta \text{educ})$$

Derive the formula of marginal effect of age on probability of being employed.

- (c) Would you agree with the following statement and why:
As a simple diagnostic one can estimate OLS in data with a binary response variable, construct a histogram of the residuals, and compare their distribution to a normal or logistic to decide whether to use probit or logit.
2. The file JTRAIN2.DAT 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. The variable *train* is the job training indicator
 - (a) Run a linear regression of *train* on several demographic and pretraining variables: *unem74*, *unem75*, *age*, *educ*, *black*, *hisp*, *married*. Are these variables jointly significant at the 5% level?

- (b) Estimate a probit version of the linear model in part (a). Compute the likelihood ratio test for joint significance of all variables. What do you conclude?
- (c) Based on your answers to parts (a) and (b), does it appear that participation in job training can be treated as exogenous for explaining 1978 unemployment status? Explain.
- (d) Run a simple OLS 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 employed in 1978? Is it statistically significant?
- (e) Run a probit of *unem78* on *train* and report the results in equation form. Does it make sense to compare the probit coefficient on *train* with the coefficient obtained from the linear model in part (d)?
- (f) Find the fitted probabilities from parts (d) and (e). Explain why they are identical.
- (g) Add all the controls from part (a) both to the linear regression and to the probit regression of *unem78* on *train*. Are the fitted probabilities now identical?
- (h) Using regressions from part (g) estimate average partial effect of all variables in linear and probit model. How do they compare?
- (i) Re-estimate regression in (g) as logit, calculate its average partial effect and compare it with your results in (h).