

Homework 4
Panel Data Analysis
Prf José Fajardo

1. Suppose an agent i is presented with two choices, to work or to be unemployed. We codify this choice by a variable $Y_i = 1$ when the agent works and 0 otherwise. The agent's utility from each choice is unobserved but it depends on an observed set of factors. To keep things simple and concrete, suppose only education, $educ$, and health, $health$, have an effect on the agent's utility. Then the agent's utility from choosing unemployment is

$$U_{0i} = \beta_{00} + \beta_{0e}educ_i + \beta_{0health}health_i + u_{0i}$$

and utility from employment is

$$U_{1i} = \beta_{10} + \beta_{1e}educ_i + \beta_{1health}health_i + u_{1i}$$

where u_{0i} and u_{1i} denote unobservable determinants of preference. Assume that u_{0i} and u_{1i} are i.i.d. $N(0, 1/2)$ and independent of each other.

- (a) Show that this defines a Probit model with dependent variable Y and exogenous variables $educ$ and $health$.
 - (b) Can you estimate the effect of education on the utility of unemployment? Can you estimate the effect of health on the difference in utility between employment and unemployment?
2. Consider the linear probability model

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where $P(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$.

- (a) Show that $E(u_i|X_i) = 0$.
 - (b) Show that $\text{var}(u_i|X_i) = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$.
 - (c) Is u_i heteroskedastic? How does this affect your inference?
3. Let $grad$ be a dummy variable for whether a student athlete at a large university graduates in five years. Let $hsGPA$ be the student's high school grade point average and SAT be the student's SAT score. Let $study$ be the number of hours spent per week in organized study hall. Suppose that, using data on 420 students-athletes, the following logit model is obtained:

$$P(grad = 1|hsGPA, SAT, study) = \Lambda(-1.17 + 0.24hsGPA + 0.00058SAT + 0.073study),$$

where Λ is the logit function. Holding $hsGPA$ fixed at 3.0 and SAT fixed at 1200, compute the estimated difference in the graduation probability for someone who spent 10 hours per week in the study hall and someone who spent five hours per week.

Repeat this exercise using Φ , the normal cumulative distribution function, instead of Λ .

4. For the model in question 1, generate the following data set:

```
set seed 123456789
set obs 50
gen educ=10+sqrt(2)*invnorm(uniform())
gen health=100*uniform()
gen u1=1.0+5.0*(educ-10)+0.1*(health-50)+sqrt(0.5)*invnorm(uniform())
gen u0=0.0+1.5*(educ-10)+0.2*(health-50)+sqrt(0.5)*invnorm(uniform())
gen Y=u1>=u0
```

Here the average number of years of education is 10 with a standard deviation of 2 and health is given as a score from 0 to 100 and is uniformly distribute for this particular population.

- (a) Interpret the data generated above. Plot the data *educ*, *health*, and *Y*, and comment on the relationships.
 - (b) Estimate three models, a linear probability model, a probit model, and a logit model for this data set.
 - (c) How much more/less likely is the average person with 12 years of education to be employed than a person with 11 years of education? Use all three models estimated above to answer this question.
5. The data file `phillipines.xls`, contains the results from a survey of small scale farmers in the Philippines. The focus of the survey is to obtain insight into the determinants of the use of chemical fertilizers.

The data set include *QFER*, the quantity of chemical fertilizers used. But we are interested in the decision to use fertilizer so we will need

```
gen FERUSE = QFER > 0
```

FERUSE is equal to 1 if fertilizer was used and 0 otherwise.

Potential determinants of fertilizer use include distance of the farm farm from the nearest market, *DMARKET*, the number of hours farmer met with an agricultural 'expert', *HOURLMEET*, farm ownership, *OWNER* (equal to 1 if farmer owns some land), farm irrigation, *IRSTAT* (equal to 1 if the farm is irrigated), and amount of loans obtained by farmer, *CREDIT*.

- (a) Estimate three models, a linear probability model, a probit model, and a logit model for this data set.
- (b) How much more/less likely is it that the average farmer will use fertilizers if he becomes an owner? Use all three models estimated above to answer this question.
- (c) Test the fit of the model using both the prediction measure and the pseudo measure.