EC220 classes

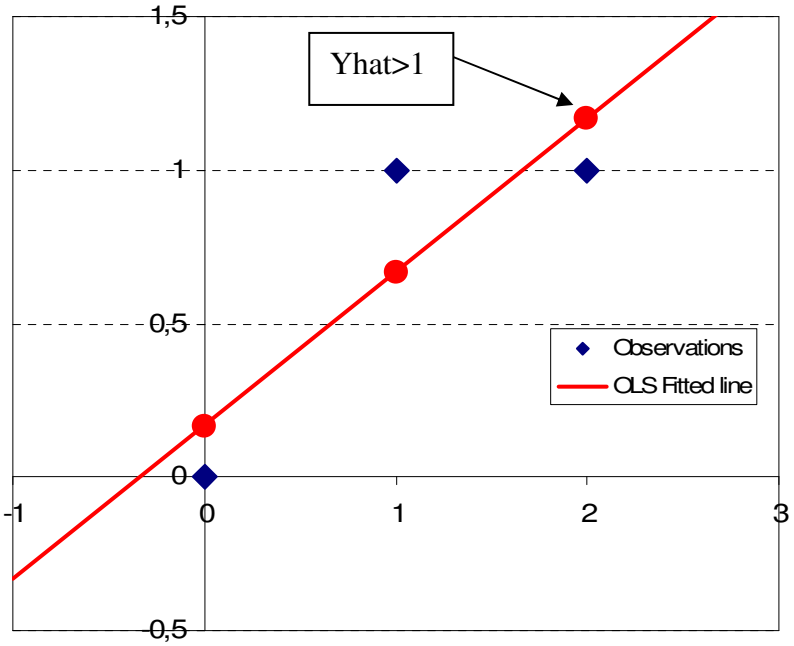
February 20, 2009

Antoine Goujard

Binary choices and Heckman selection model

1) *"As linear probability model has some defaults eg. heteroscedastic, u term not a normal distribution, so does that mean the use of logit and probit analysis can solve all these defaults? How and why these kinds of analysis can solve the problems? is it because they are continuous distribution?"*

Assume you are interested in a particular event (eg. going to college). Y_i takes value 1 if the event occurs, 0 otherwise. You want to know the effect of one explanatory variable X_i on the probability or likelihood that this event occurs.
First of all the logit and probit solves (by their definitions) the problem of the predicted values that may be either $\hat{Y}_i < 0$ or $\hat{Y}_i > 1$ in the linear probability model (LPM).
In the linear probability model (LPM fitted by OLS), we assume that the probability that $Y_i = 1$ for an individual with observable characteristics X_i is given by:
$\mathbb{P}(Y_i = 1|X_i) = \mathbb{E}(Y_i|X_i) = p_i = \beta_1 + \beta_2.X_i$.
So, after our OLS analysis (minimizing the RSS) we obtain:
$\hat{p}_i = \hat{Y}_i = b_1 + b_2.X_i$



But by construction for some value of X_i, $\hat{Y}_i$ may be greater than one or lower than zero (if you extrapolate from your sample, or even for some values in your sample, above picture).
The logit and probit do not have this disadvantage because they are based on a particular

assumption:
$\mathbb{P}(Y_i = 1|X_i) = p_i = F(Z_i)$ where Z_i is a linear function of the explanatory variables.
By definition (it is a cumulative density function, logit or normal), F is chosen such that for all possible values of Z_i, $0 \leq F(Z_i) \leq 1$. Hence the first requirements that predicted values can be interpreted as "probabilities" ($\in$ [0,1]) is satisfied.
Moreover, **the estimates for the parameters in the function Z_i do not rely on the properties on the disturbance term of the linear probability model**. They are fitted using **maximum likelihood**.
Let see how this works:
(a) If you observe $Y_i = 1$ and X_i the probability of this observation is p_i.
(b) If you observe $Y_i = 0$ and X_i the probability of this observation is $1 - p_i$.
So in general, the probability of an observation (Y_i, X_i) is:
$\mathbb{P}(Obs_i) = p_i^{Y_i}.(1 - p_i)^{(1-Y_i)}$ (You can check this using cases (a) and (b)).
The probability or **likelihood** of the sample is then given by:
$L(Y_1, ..., Y_n, X_1, ..., X_n, \beta_1, \beta_2) = \prod_{i=1}^{n} \mathbb{P}(Obs_i)$.
And you can maximize this quantity with respect to β_1, β_2 to find their maximum likelihood estimators (p313).
However now, the results are based on **asymptotic theory** ($n \to +\infty$) and you have to use asymptotic t-tests or likelihood ratio tests to test the significance of the parameters.

2) "Also, for the heckman procedure on P . 310, it said the selected variables can be used to check if they have influenced whether the dependent variable is observed? I don't quite understand the use of the select variables. Also, on P 312 it said , " if Child 06 is included in the earning function it has a positive coefficient sig at 5I don't really get this point because the coefficient of CHILD 06 in the table is -0.3982738, so why does it have a positive coefficient sig."

The Heckman selection model has two steps:
(a) The observation is observed or not (i.e. The individual decides to participate -go to college- or not).
(b) An equation of interest where the dependent variable (eg. earnings) is only observed if the individual participate.
The step (a) is modeled as the usual **probit** model where the dependent variable is participating (or selecting one-self). This is the second part of your STATA output p311 under "select". All the variables after "select" are assumed to influence the decision to participate.
Some variables may have two effects: one on the probability to participate in step (a) and an other one on the final outcome of interest in step (b). This is the case of CHILDL06 p311. The point estimate -0.398 is for the probit model, but you could also include this variable in step (b), the final equation for earnings. In this case which is not shown in the book, the estimate appears positive in step (b).