ECON 626: Empirical Microeconomics

To Probit or Not to Probit?

Department of Economics
University of Maryland
Fall 2019

1. Consider a probit data-generating process: given two independent, normally-distributed random variables x_1 and x_2, let the probability that $y = 1$ be given by:

$$\Pr[y = 1|\boldsymbol{X}] = \Phi(\beta_1 x_1 + \beta_2 x_2).$$

In other words, $\Pr[y = 1|\boldsymbol{X}] = \Pr[\beta_1 x_1 + \beta_2 x_2 > \epsilon]$ for some $\epsilon \sim \mathcal{N}(0, 1)$.

(a) Write a `.do` file that simulates this date-generating process in a sample of ten thousand observations for the parameter values: $\beta_1 = 2$, $\beta_2 = 3$, $\mu_{x_1} = \mu_{x_2} = 0$, and $\sigma^2_{x_1} = \sigma^2_{x_2} = 1$. Specifically, you should generate the following variables: x_1, x_2, y, ϵ, and $\Pr[y = 1|\boldsymbol{X}]$. Make a scatter plot of the relationship between the probit probability and $\boldsymbol{X}'\boldsymbol{\beta}$.

(b) Fit a linear probability model by regressing y on $\boldsymbol{X}$. Store the predicted values of $\widehat{y}$ as a new variable. Make a scatter plot that compares the relationship between the probit probability and $\boldsymbol{X}'\boldsymbol{\beta}$ to the relationship between $\widehat{y}$ and $\boldsymbol{X}'\boldsymbol{\beta}$. Does the linear probability model provide a reasonable fit?

(c) Notice that many of the predicted values of $\widehat{y}$ are outside the $[0, 1]$ interval. Summarize the probit probabilities for these observations.

(d) Fit a second linear probability model by regressing y on $\boldsymbol{X}$ in the interval $\boldsymbol{X}'\boldsymbol{\beta}$ such that $\Pr[y = 1|\boldsymbol{X}] > 0.05$ and $\Pr[y = 1|\boldsymbol{X}] < 0.95$. Save the predicted values of the dependent variable as $\widehat{z}$. How do the OLS coefficient estimates from your answer to (1d) compare to the OLS coefficients reported in (1b)? Make a scatter plot comparing the probit probabilities and the predicted values of $\widehat{z}$ over the range of values of $\boldsymbol{X}'\boldsymbol{\beta}$ such that $\Pr[y = 1|\boldsymbol{X}] > 0.05$ and $\Pr[y = 1|\boldsymbol{X}] < 0.95$. Does the linear probability model provide a reasonable fit *within this interval*?

2. Now consider an alternative data generating process: let

$$\Pr[y = 1|\boldsymbol{X}] = \Pr[\beta_1 x_1 + \beta_2 x_2 > \eta]$$

where η is defined as

$$\eta = \begin{cases} \zeta_1 & \text{if } \mu > 0 \\ \zeta_2 & \text{if } \mu < 0 \end{cases}$$

for independent random variables $\zeta_1 \sim \mathcal{N}(-4, 1)$, $\zeta_2 \sim \mathcal{N}(4, 1)$, and $\mu \sim \mathcal{N}(0, 1)$.

(a) Write a `.do` file that simulates this data-generating process in a sample of one hundred thousand observations for the parameter values: $\beta_1 = 2$ and $\beta_2 = 3$. Make a histogram of η. Use the `cumul` command to generate the empirical CDF of η, and present it as a line graph.

(b) Use the `lpoly` command to plot a locally-weighted non-parametric kernel regression of the probability that $y = 1$ as a function of $\boldsymbol{X}'\boldsymbol{\beta}$.[1]

(c) Fit probit and OLS models of $y = 1$ as a function of x_1 and x_2. Store the predicted probabilities and graph these together with the kernel regression probabilities as functions of $\boldsymbol{X}'\boldsymbol{\beta}$. How does the fit of the linear probability model compare to the fit of the non-linear probit model?

[1] A lower-tech approach that would yield more or less the same result would be to calculate an empirical estimate of the probability that $y = 1$ by sorting the data by $\boldsymbol{X}'\boldsymbol{\beta}$ and dividing it into bins of one thousand observations each. You could then calculate the empirical probability that $y = 1$ within each bin.