ECONOMICS 717
APPLIED ECONOMETRICS
Department of Economics
University of Wisconsin-Madison
Spring 2022
Problem Set: Binary Dependent Variable Models
Version of February 4, 2022
Due on Canvas at 11:00 AM on Wednesday, February 16, 2022

Introduction

This problem set uses data from Field, Jayachandran, and Pande (2010) *American Economic Review*. The paper studies the effect of randomly assigned entrepreneurship courses for poor, self-employed women in India. The primary outcome variable in the paper, and in the problem set, is loan take-up in the post-random-assignment period. The paper is available on the CTools site if you want to learn more about the underlying analysis.

Answers

Your write-ups for the problem sets should consist of two portions. The first portion is just the answers to the questions, with whatever text is required to explain them. The second portion, on separate pages, consists of a Stata log file that shows how you got the answers to the empirical questions. The log file must be clear and must include comments that will allow the grader to quickly see the command or commands leading to each answer. It should not include everything you tried – just the final set of commands employed to get the answers. See the syllabus for more details on the format.

Data set: analysis file from Field et al. (2010)

The data set is called "Field et al. (2010) Analysis File" and can be found on Canvas page.

The variables used in the problem set are:

imidlineid: ID number
taken_new: 1/0 the client took out a loan in the four months before the midline survey
Client_Age: age of the client in years
Client_Married: 1/0 the client is married
Client_Education: years of schooling of the client in years
HH_Size: number of persons in the client's household
HH_Income: client's household income
muslim: 1/0 the client is Muslim
Hindu_SC_Kat: 1/0 the client is a Hindu and in a scheduled caste
Treated: 1/0 the client was random assigned to the experimental treatment group
miss_Client_Age: 1/0 missing value for client's age
miss_Client_Married: 1/0 missing value for client's marital status
miss_Client_Education: 1/0 missing value for client's education

Problems

1. Drop observations with missing values of client age or client marital status or client education or client household income. Do not use the *miss_HH_income* variable, which is incorrectly coded in the data. Instead, drop the observations with missing values directly. The missing indicators for client age, education and marital status are correctly coded and so may be used for this problem.

2. Estimate a linear probability model with loan take-up as the dependent variable and client age, client marital status, client years of schooling, client household size, client household income, indicators for Muslim and Hindu scheduled caste, and experimental treatment status. Estimate the model without the `robust` option. Report and discuss the coefficient estimates.

3. Repeat the estimation in Problem 2 with the `robust` option. How big are the differences, if any? Which standard errors are larger?

4. Generate predicted probabilities of loan take-up using the estimated coefficients from Problem 2. Do any of these probabilities lie outside [0, 1]? If so, do the observations corresponding to these values show any particular patterns in the values of the variables?

5. Estimate the model by weighted least squares using Stata's `vwls` (variance weighted least squares) command. How much do the coefficients differ from those obtained without weighting? How much do the standard errors differ from those produced by the `robust` option?

6. Estimate probit and logit models of loan take-up using the same independent variables as in Problem 2. Are the probit and logit coefficient estimates similar to one another and to the LPM estimates? Should they be? Explain why or why not.

7. Calculate the mean partial derivatives (a.k.a. marginal effects or average partial effects) of the conditional probabilities of loan take-up with respect to client age for the LPM, logit and probit models. Explain in words what these derivatives mean.

For the probit model, calculate the derivatives using four different methods:

(a) Calculate analytic derivatives evaluated at the mean of the covariates. Stata will do this using the `dprobit` command.

(b) Calculate mean analytic derivatives by hand. That is, use the formula to calculate the partial derivative for each observation and then `summarize` these to get the mean.

(c) Calculate mean numerical derivatives by calculating predicted probabilities for each observation and then changing the client age value for each observation by a little bit (e.g. 0.001) and recalculating the predicted probabilities.

(d) Calculate mean derivatives using the `margins` command.

For the probit model, do the four methods of calculating the partial derivatives give about the same answer?

Do the three models - LPM, logit and probit - yield similar estimates of the mean partial derivatives?

8. Re-estimate the LPM including a quartic in client age (i.e. including linear, squared, cubed and fourth power terms), rather than just a linear term. Calculate the average derivative numerically, as in part (c), and compare it to that obtained from the probit model. Does the LPM do better with greater flexibility?

9.. For the probit model estimated in Problem 6, calculate the value of the LRI "by hand" by obtaining the likelihood values for the model in Problem 6 and a model with only an intercept and plugging them into the LRI formula. Interpret the value you obtain.
.
10. Calculate the correct prediction rate for the probit model estimated in Problem 6 using both 0.5 and the sample fraction taking up a loan as cutoff values. In each case, take the equally weighted average of the correct prediction rates for those who take up a loan and those who do not. Indicate and discuss how the model performs in each case. Explain any differences between the two and indicate which measure of goodness of fit you prefer and why. [Note: the equal weights refer to the prediction rates, not the individual observations.]

11. Examine the difference between in-sample and out-of-sample predictive success by estimating the model using the usual covariates but only those observations with *imidlineid* < 1400. Compare the predictive performance of the model using the same criteria as in the preceding question for observations in and out of the estimation sample. What differences, if any, do you find? Does what you find accord with your expectations? Explain.

12. Estimate a probit model of loan take-up including the usual covariates plus an interaction term between married and Muslim.

13. Compare mean finite differences of the interaction term calculated without the additional terms highlighted in the Ai and Norton (2003) paper in the probit model estimated in Question 12 with mean interaction effects that included these terms calculated "by hand" in Stata Are they very different? Be sure to compare finite differences rather than derivatives.

14. What is the standard deviation of the estimated interaction effects obtained "by hand" in Problem 13 across sample observations? Why is this variance so small in this context?

15. Estimate a regression of the squared residuals from a linear probability model of loan take-up with the usual covariates on the usual covariates. What evidence, if any, do you find of heteroskedacity?

16. Using the `hetprob` command, estimate a probit model of loan take-up on the usual covariates, allowing for heteroskedasticity that depends on client age and education. What evidence, if any, do you find of heteroskedasticity?