DATA DESCRIPTION

BINARY ENDOGENOUS TREATMENT IN STOCHASTIC FRONTIER MODELS WITH AN APPLICATION TO SOIL CONSERVATION IN EL SALVADOR

SAMUELE CENTORRINO,MARÍA PÉREZ URDIALES,BORIS BRAVO-URETA,ALAN WALL

1. Data source

Data are from *Programa Ambiental de El Salvador* or PAES, an environmental program promoting crop diversification and soil conservation practices. The data set was obtained by one of the co-authors (Boris Bravo-Ureta), who participated in and oversaw the implementation of this program. The total sample size is 459.

2. List of variables

The database *elsavadordata.txt* contains the cross-section of farmers sampled in 2005. For each farmer, the following variables are available:

- *louput*: log of the total value of production, measured in dollars.
- *lland*: log to the total cultivated area (in manzanas where one manzana=0.7 has).
- *llabour*: log of the total number of workers (family + hired).
- *lseeds*: log of the value of seeds used (measured in dollars).
- *lfertilizers*: log of the value of fertilizers used (measured in dollars).
- *lpesticides*: log of the value of pesticides used (measured in dollars).
- *Participation*: dummy variable for participation into the soil conservation program. 1 if farmer has participated, and 0 otherwise.
- *fullten*: dummy equal to 1 if the farmer owns all her cultivated land, and 0 otherwise.
- *ageold60*: dummy equal to 1 if the farmer is 60 and older and 0 otherwise.
- *maxeduc*: highest level of education in the farmer's household (measured in years).
- *nooutinc*: dummy equal to 1 if the farmer has no outside sources of income and 0 otherwise.
- *footaccess*: dummy, equal to 1 that indicates whether farmers have only walking access to their plot, and 0 otherwise.
- *caraccess*: dummy equal to 1 that indicates whether farmers also have access by car to their plot, and 0 otherwise.

Date: October 10, 2023.

- *riskmax*: defined as a continuous variable which compares the relative diversification of each farmer with respect to the average farmer. A higher value of the *riskmax* index implies higher risk diversification.
- *valorjornal*: average daily wage within the 'canton'.
- *fam_electricidad*: proportion of families living in the same 'canton' with access to electricity.
- *fam_letrinas*: proportion of families in the farmer's canton with access to private bathrooms.
- *diste1*: log-distance (measured in kilometers) from the epicenter of the earthquake in 2001.
- *dum01*: Dummy for region 1.
- *dum02*: Dummy for region 2.
- *dum03*: Dummy for region 3.
- *dum04*: Dummy for region 4.

3. Matlab Code

A Matlab code which is used to obtain the results of the paper is provided with the data. The following scripts allow the reader to reproduce the Monte-Carlo simulations and the empirical application.

- *cpuw_montecarlo.m*: contains the code for the main DGP used in Section 3.
- *run_mc_cpuw.m*: runs the previous script for each sample size $n = \{250, 500, 1000\}$, and value of the dependence parameter $\rho_U = \{0, 0.5, 0.95\}$.
- *cpuw_empirical.m*: main script for the empirical application. Reproduces estimation for the exogenous and the endogenous model with tests and confidence intervals.
- *liksf.m*: Likelihood function for the normal-half-normal stochastic frontier model with scaled exponential inefficiency.
- *likendsftreat.m*: likelihood function for the normal-folded-normal stochastic frontier model with scaled exponential inefficiency, where the treatment dummy is allowed to be endogenous.
- *teex.m*: function to compute the efficiency scores in the normal-half-normal model stochastic frontier model with scaled exponential inefficiency.
- teendbin.m: function to compute the efficiency scores in the normal-folded-normal model stochastic frontier model with scaled exponential inefficiency.