ECON-665  
Homework 3

260535788 Kun Yang

1. **Simple DD**

*a)* *Find some summary statistics:*



The number of observations is 1652.



There are 826 households in this dataset.



The mean of total expenditures overall is 4858.965.

The mean of total expenditures in 1991 is 4163.101

The mean of total expenditures in 1998 is 5554.829.

*b) Explain how the commands worked*

Create a temporary variable exptot0, which contains the 1991 total expenditure.

**gen exptot0=exptot if year==0**

Create a new variable exptot91, which includes only the values of 1991 expenditure from temporary variable exptot0. Make these values also available for the same household in 1998 by option [**by(nh)**]. Thus, the 1991 expenditure is available in observations of both 1991 and 1998.

**egen exptot91=max(exptot0), by(nh)**

As the 1991 expenditure were stored in the 1998 observations as a new variable exptot91, we can delete all 1991 observations and keep only 1998 observations.

**keep if year==1**

Create a 1991 level log per capita expenditure variable.

**gen lexptot91=ln(1+exptot91)**

Create a 1998 level log per capita expenditure variable.

**gen lexptot98=ln(1+exptot)**

Calculate the differences between 1998 and 1991 per capita expenditures (in log) for every observations and save it as lexptot9891.

**gen lexptot9891=lexptot98-lexptot91**

*c) Take the difference variable of outcomes created earlier (“lexptot9891”) and compare it for microcredit participants and nonparticipants.*

****

**T-test**

**Null hypothesis:** the mean of the differences between 1998 and 1991 per capita expenditures for *the households with female participants* is equal to the mean of the differences between 1998 and 1991 per capita expenditures for *the households without nonparticipants.*

The p-value of the t-test is 0.0021 (<0.05). Thus, we can reject the null hypothesis at 95% confidence interval.

At 95% confidence level, we can conclude that there is a statistically significant difference between the mean of the 1998-1991 per capita expenditures differences for the households with female participants and the households without female participants.

**2. Ordinary Least Squares Regression**

*a) Run OLS*

**

The independent variable, log total expenditure per capita (lexptot), is the outcome of interest. The year dummy equals 1 for 1998 (the second period), and it captures the time trend even without the program. The treatment variable dfmfd98 equals 1 for the treatment group, and it captures differences between the treatment and control groups prior to the program. The interaction term (dfmfdyr) is same as a dummy variable equals to 1 for those observations in the treatment group in 1998. It captures the impact of the microcredit program participation on total expenditure per capita. The coefficient of interest is the coefficient of the interaction term.

The estimated coefficient of the interaction term (dfmfdyr) is 0.111 (P-value=0.014<0.05). It is statistically significant at 95% confidence level. It means that microcredit program can increase the participants’ total expenditure per capita by 11.1% on average from 1991 to 1998, ceteris paribus.

The estimated coefficient of the treatment variable (dfmf98) is -0.115 (P-value=0.00<0.05). It is statistically significant at 95% confidence level. It means that, in the year of 1991, 1998 participants (treatment group) are lower in expenditure by 11.5% on average than the 1998 nonparticipants (control group), ceteris paribus.

The estimated coefficient of the year dummy is 0.147 (P-value=0.00<0.05). It is statistically significant at 95% confidence level. It means that the common time trends exist, and even the nonparticipants’ expenditure increased by 14.7% from 1991 to 1998, ceteris paribus.

The adjusted R-squared equals to 0.0536. This OLS regression is very poorly fitted.

*b) Run the regression model by including other covariates that may affect the outcomes of interest:*



There are some other covariates that may change over time and have effect on the expenditure. So, we need to add more covariates in the regression to control their effect.

After controlling additional covariates, the estimated coefficient of the interaction term (dfmfdyr) is 0.051 (P-value=0.332>0.05). It is **NOT** statistically significant at 95% confidence level. It indicates that microcredit program have no impact on the participants’ total expenditure per capita after controlling additional covariates, at 95% confidence level.

The estimated coefficient of the treatment variable (dfmf98) is 0.001 (P-value=0.970>0.05). It is **NOT** statistically significant at 95% confidence level. It indicates that, in the year of 1991, the 1998 participants (treatment group) are not different from the 1998 nonparticipants (control group) on the expenditures.

The estimated coefficient of the year dummy is 0.277 (P-value=0.00<0.05). It is statistically significant at 95% confidence level. It means that the time trends exist, and even the nonparticipants’ expenditure increased by 27.7% from 1991 to 1998, ceteris paribus.

For other variables, the estimated coefficients of educhead, lnland, milk and egg are statistically significant at 95% confidence level. It indicates that these variables also have impact on the total expenditure per capita. All other additional covariates are NOT statistically significant.

The adjusted R-squared increased to 0.2826. This regression with additional covariates fitted better than the one in part (a) but still not good enough.

**3. Fixed-effects regression**

*a) Fixed effects regression controls for household’s unobserved and time-invariant characteristics that may influence the outcome variable.*



The estimated coefficient of the dfmfdyr is 0.111 (P-value=0.002<0.05). It is statistically significant at 95% confidence level. It means that microcredit program can increase the participants’ total expenditure per capita by 11.1% on average from 1991 to 1998, ceteris paribus.

The estimated coefficient of the treatment variable (dfmf98) is omitted because we are using fixed-effect in this program and the time-invariant unobservable characteristics of the households have been controlled already.

The estimated coefficient of the year dummy is 0.147 (P-value=0.00<0.05). It is statistically significant at 95% confidence level. It means that even the nonparticipants’ expenditure increased by 14.7% from 1991 to 1998, ceteris paribus.

The adjusted R-squared equals to 0.1450. This OLS regression is poorly fitted.

*b) Regress the same outcome against the village program indicator variable, plus other factors that may influence the expenditure.*

**

There are some other covariates that may change over time and have effect on the expenditure. So, we need to add more covariates in the regression to control their effect.

After controlling additional covariates, the estimated coefficient of the dfmfdyr is 0.091 (P-value=0.014<0.05). It is statistically significant at 95% confidence level. It indicates that, after controlling additional covariates, microcredit program can increase the participants’ total expenditure per capita by 9.1% on average from 1991 to 1998, ceteris paribus.

The estimated coefficient of the treatment variable (dfmf98) is omitted because we are using fixed-effect in this program and the time-invariant unobservable characteristics of the households have been controlled already.

The estimated coefficient of the year dummy is 0.221 (P-value=0.000<0.05). It is statistically significant at 95% confidence level. It means that, after controlling additional covariates, even the nonparticipants’ expenditure increased by 22.1% on average from 1991 to 1998, ceteris paribus.

For other variables, only the estimated coefficients of inland and egg are statistically significant at 95% confidence level. It indicates that the village price of egg and households’ land size also have impact on the total expenditure per capita. All other additional covariates are not statistically significant.

The adjusted R-squared increased to 0.1715. This regression fitted a little bit better than the previous one but not improved a lot.

*c) Could you do other estimations in order to estimate the impact of the microcredit program? Which ones? What test(s), if any, would you run? Would you modify the commands above?*

**Method 1)** We can redo the t-test, but compare the mean of participant-nonparticipants difference between 1991 and 1998.

Create a temporary variable exptot0, which contains the total expenditure for nonparticipants

**gen exptot0=exptot if dfmfd==0**

Create a new variable exptotout, which includes only the values of nonparticipant expenditure from temporary variable exptot0. Make these values also available for the same household by option [**by(nh)**]. Thus, the nonparticipant expenditure is available in observations of both participants and nonparticipants.

**egen exptotout=max(exptot0), by(nh)**

As the nonparticipant expenditure were stored in the participant observations as a new variable exptotout, we can delete all nonparticipant observations and keep only participant observations.

**keep if dfmfd==1**

Create a nonparticipant log per capita expenditure variable.

**gen lexptotout=ln(1+exptotout)**

Create a participant log per capita expenditure variable.

**gen lexptotin=ln(1+exptot)**

Calculate the differences between participant and nonparticipant per capita expenditures (in log) for every observations and save it as lexptotinout.

**gen lexptotinout=lexptotin-lexptotout**

**

**T-test**

**Null hypothesis:** the mean of the differences between participant and nonparticipant per capita expenditures in 1991 is equal to the mean of the differences between participant and nonparticipant per capita expenditures in 1998.

The p-value of the t-test is 0.0001 (<0.05). Thus, we can reject the null hypothesis at 95% confidence interval.

At 95% confidence level, we can conclude that there is a statistically significant difference between the mean of the participant-nonparticipant per capita expenditures differences for the households in 1991 and the households in 1998.

**Method 2)** We can also calculate the 1991-1998 difference first. Then we can take the difference between 1991-1998 difference in the treatment and comparison Groups.

## Transform the dataset into wide form.

**keep lexptot dfmfd nh year sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg**

**reshape wide lexptot dfmfd sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg, i(nh) j(year)**

## Generate the difference between 1991 and 1998

**gen dy = lexptot1 - lexptot0**

**replace dfmfd0=0**

**gen dpdfmfd = dfmfd1-dfmfd0**

**gen dpsexhead = sexhead1-sexhead0**

**gen dpagehead = agehead1-agehead0**

**gen dpeduchead = educhead1-educhead0**

**gen dplnland = lnland1-lnland0**

**gen dpvaccess = vaccess1-vaccess0**

**gen dppcirr = pcirr1-pcirr0**

**gen dprice = rice1-rice0**

**gen dpwheat = wheat1-wheat0**

**gen dpmilk = milk1-milk0**

**gen dpoil = oil1-oil0**

**gen dpegg = egg1-egg0**

## do the regression of difference in difference.

**reg dy dpdfmfd**

**reg dy dpdfmfd dpsexhead dpagehead dpeduchead dplnland dpvaccess dppcirr dprice dpwheat dpmilk dpoil dpegg**



This method give us the **same results with the fixed effect model**.

The estimated coefficient of the dpdfmfd is 0.111 (P-value=0.002<0.05). It is statistically significant at 95% confidence level. It means that microcredit program can increase the participants’ total expenditure per capita by 11.1% on average from 1991 to 1998, ceteris paribus.



After controlling additional covariates, the estimated coefficient of the dpdfmfd is 0.091 (P-value=0.014<0.05). It is statistically significant at 95% confidence level. It indicates that, after controlling additional covariates, microcredit program can increase the participants’ total expenditure per capita by 9.1% on average from 1991 to 1998, ceteris paribus.

**Question 4. Assessment of results**

*What do you think one can say about the impact of the microcredit program on household expenditures?* *Should one trust the results using the DD methodology? Why? Or why not?* *What type of information would you need in order to assess further the validity of the DD methodology to assess the impact of the program in this particular case?*

The DD methodology cannot be trusted. The DD methodology (question 2) and simple t-test (question 1) assumes that there exist a common trend across time, but we cannot make sure such common trend exist. It could be the case that households in different locations experienced different GDP growth rate. The different economic growth rate can make our estimation biased.

We need the information that before 1991. Then, we could compare the expenditure trend or some other village characteristics in these different locations before 1991. If the common trend exist before the initial survey, then maybe the assumption of common trend cross time exist.

The fixed effect model generalize and improve the DD methodology, and do not assume such parallel common trend exist. It captures the unobservable time-invariant effect. Thus, the fixed effect model with additional covariates is more reliable than the simple t-test and OLS DD methodology.

We can say that the impact of the microcredit program is positive 9.1% on household expenditures and it is significant. But the R-squared is only 0.1715, it could be a worry.