UNIUNEA EUROPEANĂ

GUVERNUL ROMÂNIEI

Instrumente Structurale
2007 - 2013

Counterfactual Impact Assessment

Day 4

Klara Major

PROJECT COFINANCED BY EFRD THROUGH TAOP 2007-2013

Regression techniques in estimating the counterfactual effects

- Ignorability of treatment
- Difference in difference regression models with control variables
- Difference in difference regression models with fitted propensity score variable

.

Econometric techniques estimating counterfactual effects

- So far we have seen
  - Pooled cross-section and panel data models basically used dummy variables for the treated. We focused on the difference in the averages of treated and non-treated
  - In counterfactual settings there is a self-selection bias that must be eliminate
  - Matching techniques eliminate the self-selection bias by the special construction of the control group
- Regression models aim to eliminate the self-selection bias by introducing control variables into the equation

- ATE = average treatment effect
  - This measures the average effect for the whole population
- ATE1 = average treatment effect on the treated
  - This measures the average effect only for the treated subpopulation

- Generally

  ATE1 = ATE + person specific gain from participation

  - person-specific gains are the gains **over** the average gain
  - eg. If we can assume that the effect of the program is the same for each participant then person specific gains are zero

Case of randomization

- Notation:
  - w binary variables, indicates participation in the program
  - y outcome, variables of interest
- Assumption
  - w is statistically independent of y
- Under this condition
  - ATE = ATE1
  - Diff-in-diff estimator is appropriate to estimate the average effect of the treatment
  - These results also hold if we assume a bit less than randomization

Mean indepedence condition

- Mean independe:
- the average outcome for the treated and non-treated groups are the same
  - Before the treatment and after the treatment the averages of the different subsamples are both the same
- Usually this is not the case

- Participation is a decision of individuals that probably depends upon the expectable outcome of the program
  - There is a self-selection into the program

- ATE1 can be estimated as a simple difference in means under weaker conditions as well
- Assume that pre-treatment outcome is independent from the participation in the program
  - Averages of the treated and non-treated are the same before the program
- However this is still a too strong assumption
  - Basically it sais that participation decision is not related to the potential future gain from the program

Ignorability of Treatment

- Assume, that mean independence is possible to achieve after controlling for the difference between individual
  - **Conditional mean independence** assumption
- This is the so-called **Ignorability of Treatment** assumption
- We need **control variables** to discover the factors that lead to the different gains for different subgroups of population
- After controlling for these variables we can estimate the effect of treatment
  - By simply calculating the difference in averages in the sample (diff-in-diff)
- This is a very strong assumption
  - Selection on observables
  - Other methods (IV, panel) do not apply this assumption

Regression methods

- Simplest methods assume the ignorability of treatment conditions
- We have two classes to estimate:
- Regression methods
  - Use control variables directly in the regression equation
  - variant of diff-in-diff estimator
- Propensity score models
  - Use control variables to estimate the chance of being treated, that is it estimates a propensity score
  - Apply the fitted value of the propensity score in a diff-in-diff equation to control for the self-selection bias

- These estimation thechniques only work if we have treated and non-treated for the different values of the control variables
- Example
  - Assume we have only 1 control variable: pre-treatment income. X is a binary variable: 1 if below a certain treshold, otherwise 0
  - Suppose everyone in the x=1 (relevant) population participates in the program
  - In this case we can not estimate what will be average effect of the treatment in this low pre-treatment income level group
  - Conclusion is similar if we observe that no one in the high income level group participates in the program
- Conclusion: only in that range of the contol variables possible to estimate the effect of the program where both treated and non-treated can be found in the sample

In practice

- Those subgroups that are not eligible for the program must be left out from the analysis
- Those subgroups from which every unit participates in the program must be left out of the analysis
  - Alternatively: we need to choose different methods for estimating the effect, IV or panel

Diff-in-diff regression: Case 1

- Assume that the predicted person-specific gain of the program is zero in every subpopulation
- As we seen earlier in this case ATE = ATE1
- The general (average) effect of the program is additive and can easily be estimated by adding a binary treatment variable to the regression
- ATE is the estimated coefficient on w

- Regression: y on constant, w, x

Diff-in-diff regression: Case 2

- If we can not assume that person-specific gains are different from zero (as is generally the case) then we need to apply a different regression form for estimating the average effect of the program
- Regression: y on constant, w, x, w(x-μ)

  where μ is the sample average of x
- The estimated coefficient on w is β, and on the interaction term it is δ, then
  - estimated ATE is β
  - Estimated ATE in different subgroups of population ie, at different level of the control variables is β+δ(x-μ)

- Effects of Enterprise Zones on Economic Development
- Daniele Bondonio: Evaluating the Employment Impact of Business Incentive Program in EU Disadvantaged Areas. A case from Northern Italy, 2002
- Business incentive programs as regional economic development tools to promote employment growth in areas with severely distressed and/or declining socio-economic conditions.
- Show that Piedmont business incentive program did not significantly affect employment in the “Objective 2 areas”.

Table 2: 1994-98 employment growth in industrial SMEs

Sectors	*Geographic areas*	*1994 employment stock in industrial SMEs*	*1994-98 absolute employment change*	*1994-98 % employment change*	*Employment change difference between Obj.2 and non-Obj.2 areas (% points values)*
Target sectors[a]	Obj.2 areas*	156,83	14,905	0.095	0.024
	Non-Obj.2 areas	161,322	11,519	0.071	
Non-target sectors	Obj.2 areas*	109,706	12,824	0.116	0.017
	Non-Obj.2 areas	104,223	10,327	0.099	

* Turin Province
(a) Target sectors are those in which at least one firms received incentive payments in the 1994-98 period

Table 6: Employment impact of Obj.2-area incentive payments to SMEs
Results from eq. (1) and (3) [Dependent variable: 1994-98 employment growth]

Independent variables		Regression coefficient estimates(+) Specific. (I)	Specific. (II)	Specific. (III)	Specific. (IV)
CONSTANT		**0.166***** **(0.028)** ***0.000***	**0.161***** **(0.031)** ***0.000***	**0.166***** **(0.028)** ***0.000***	**0.159***** **(0.028)** ***0.000***
TREATMENT VARIABLE					
1 if sector and province are targets of the intervention. 0 otherwise	T	0.034 (0.077) *0.654*	0.007 (0.108) *0.947*	-	-
Fin=thousand of Euros per employee (if sector and province are targets)	Fin	-	-	0.007 (0.045) *0,929*	-0.021 (0.052) *0,826*
CONTROL VARIABLES					
Sector-specific economic trend	S	**-0.105**** **(0.042)** ***0.014***	**-0.100**** **(0.045)** ***0.026***	**-0.100**** **(0.041)** ***0.017***	**-0.097**** **(0.042)** ***0.023***
Area-specific economic trend P=1 if province is Obj.2 area 0 if province is not Obj.2 area	P	-	0.027 (0.075) *0.713*	-	0.037 (0.062) *0.542*
Number of observations		266	266	266	266
Adjusted R2		0.255	0.202	0.177	0.154
F		3.88	3.19	2.98	2.11
Prob>F		0.0176	0.0234	0.0324	0.0597

* p-value<0.1 ** p-value<0.05 *** p-value<0.01

Standard deviations in parenthesis. P-values in *italics*

Propensity score models: Case 1

- Assume that the chance of being treated is between 0 and 1
- Assume that the change in average outcome in the different control variable groups is uncorrelated with the **variance** of the participation in these groups

- Regression y on constant, w, p
  - Where p is the fitted value of the propensity score

Propensity score models: Case 2

- Assume that the chance of being treated is between 0 and 1

- Regression y on constant, w, p, w(p-π)
  - where p is the fitted propensity score
  - and π is the sample average of the fitted propensity score

- The data in JTRAIN2.GDT are from a job training experiment in the 1970s. The response variable is real earnings in 1978, measured in thousands of dollars. Real earnings are zero for men who did not work during the year. Training began up to two years prior to 1978.
- The factors influcencing income are real earnings in 1974 and 1975, age (in quadratic form), a binary high school degree indicator (nodegree), marital status, and binary variables for black and Hispanic.

Exercise

- Estimate the probability of training using probit different probability models (linear model and probit).
- Forecast these models to have the propensity score.
- Estimate average treatment effect in a diff-in-diff regression using only the treatment dummy and the fitted propensity score value (case 1). Interpret the results.
- Include the interaction term into the regression (case 2) and interpret the results.

Preparing a TOR for counterfactuals

- Decisions to make

.

Main questions

1. Expected impact of the program
2. Data background
3. Expected methodology
4. Circumstances of the analysis (time, money, etc)

Questions concerning the program

- What are the relevant impacts the program is expected to have?
- Possible answers:
  - The unemployment will decrease in the targeted low-income regions
  - Unemployed will easier find a jog
  - Increase the labor market participation of mothers with young child
  - Increase in the sales of SME-s

Questions concerning the program

- What is the relevant target group of the intervention?
- Possible answers:
  - Enterprise zones in low-income regions
  - Long-term unemployed in low-income regions, long-term means more than 9 months
  - Mothers with children where the age of the youngest child is below 6 if number of children is more than or equal to 3, or 3 if the number of children is less than or equal to 2

Questions concerning the program

- What are the relevant tools to acheve these goals?
- Possible answers:
  - Supporting enterprise zones’ infrastructure investment
  - Training of unemployed either to have a new professon or the techniques of job searching
  - Developing nursary school in quantity and quality
  - Supporting SME’ investment in tangible assets

Questions concerning the program

- What is (are) the assumed theoretical causality in the program goal and tool variables?
- Possible answers:
  - Enterprise zones offers better infrastructural background for firms therefore by supporting them they will be more attractive to firms. If number of firms in EZs' located in the low income regions increases then there might be an increase in employment in those regions
  - By having more demanded profession unemployed people will find job easily
  - By making easier to solve the problem of child-caring problems for mothers it is less costly for them to accept a job
  - By having more effective assets they can produce more and by selling the increased quantity of products (goods and/or services) they can have higher sales revenue

- Based on the answers to the previous three questions we need to express the expected impact of the program in terms of hypothesises
- Possible solutions:
  - Due to the support there is more firms in the supported EZ than in that of not supported. The overall employment of the firms in the supported EZ increased better, than in the not supported EZ's. The unemployment rate in the low-income regions where there are more supported EZ has increased better than in those low-income regions where there is lower number of EZ's.
- Exercise:
  - Draft the hypthesis in the other 3 example.

Questions in TOR

- These questions are usually answered already
  - In the preparation document of the program (before the decision have made)
  - In the interim evaluation documentation
- However it is the occasion when these chain of reasoning must be translated to the language of counterfactuals first in terms of hypothesis

Data background

- The key: we need data
  - before the program and after the program is over,
  - both for the beneficiary and the non-beneficiary (control group)
- Possible data sources:
  - Program documentation (application form, interim reports) but these are usually do not contain information on the control group!
  - Statistical office
  - Tax office
  - Institutions of the social security systems
- Data collection in the period of evaluation:
  - Usually by questionnaire
  - It is possible to ask the beneficiary and the non-beneficiary as well
  - It will not provide information on past data, especially before the program period

Expected methodolgy

- Difference-in-difference
- Propensity score models
- Instrumental variables
- Discountinuity design
- Panel data methods

- A priori it is hard to decide which one is the best
- Probably it should be left for the best judgement of the expert

Other issues

- Time requirement of the evaluation depends upon
  - How many different hypothesis must be tested
  - Availability of the data
    - Are they ready for analysis, that is collected, cleaned and organized in a dataset?
    - If not, what other works are needed to do in order to reach this state?
  - If other than counterfactual methods are expected (like soft methods, more sociologyst approaches based on surveys) then are these tasks possible to do at the same time, etc.

Preparing a TOR

- Form groups of 2-3 persons and prepare a TOR in your operating program.
- Make the decisions concerning the above specified questions.
- Write a TOR (or an outline of a TOR) based on these specifications.

Understanding a TOR

- You are an expert in a policy evaluation project.
- Your task is to plan the counterfactual estimation strategies.
- Read the TOR given to you and try to understand it.
- Based on you understanding you are allowed to put on questions once.
- Based on the information you got prepare a plan for the estimation.

Thank you for your attention!

Bucharest 30.10.2011