The reform of Administration Approval System and Firms Innovation——An application of Double Machine Learning

Xia Zou
27720171152673,WISE

May 16, 2019

1 Introduction

As China became a member of World Trade Organization on 11 December 2001 and the release of 'Administration Liscening Law', the reform of administrative approval was carried out nationally. The reform of administrative approval system required establishing administrative approval center, where government departments with approval authority located concentrationally. In this way, government are able to provide firms with registration, investment, paying national tax and local tax service conveniently(all in one place), which will enhance cooperation among different administrative approval departments and increase approval efficiency. Such regulatory policies are believed to have impact on the economy and innovation ect. . (see Blind (2012))

Wang and Feng (2018) combine data from the Quasi-Natural Experiment of establishing administrative approval center with Chinese Micro enterprise data to studies the impacts of administrative approval system reform on enterprise innovations with Difference in Difference Model and Triple difference Model. They found that administrative approval center(APC) significantly improve the level of enterprise innovation. However, as shown in Figure 1, the average number of applied patent in treated group and control group doesn't satisfy common trend assumption as required by DID model, which is not the same as stated by Wang and Feng (2018). What's more, average applied patent for treated group seems to be higher than control group before 2002 , which provides evidence for the non-randomness of treatment assignment (Whether the city that a firm located have established APC or not).

Without common trend assumption hold, it's reasonable to include control variables as Wang and Feng (2018) did in their DID model. However, does such linear specifications are able to adequately control for the number of applied patent ? On the other hand, the difficulty for estimation of treatment effect increase as we allow a more flexible model. In such case, Double Machine Learning proposed by Chernozhukov et al. (2018) provide us one resolution. Thus this paper use the same data set as Wang and Feng (2018) to estimate the impact of the reform of administrative approval system on firms' Innovation activities with Double Machine Learning(DML) method.

The remaining of this paper is arranged as follows. Section 2 includes description of data. Section 3 introduce Double Machine Learning model used in this paper as well as the corresponding algorithm. Section 4 displays results and some explanations.

2 Data and Descriptive Analysis

As mentioned before, this paper use the same dataset as Wang and Feng (2018), which can be found in ***China's Industrial Economics*** website. This dataset include Chinese enterprise micro data from 1998 to 2006. And treat group include all enterprises in cities that have established APC in 2002 while control group are all firms in cities that either established APC after 2007 or never established APC. The followings are variables included in dataset.

1. Response variable Y is defined as Ln(patent), where patent is the amount of applied patent for a firm in a specific year.

2. Policy variable D. D is defined as the product of treat dummy variable and time dummy variable. Treat is an indicator for whether the city that a firms located have established APC in 2002 or not. And time dummy indicate whether the year is before 2002 or not.

3. Control variables X. In the first specification, control variable X include age , age square, export, number of employees, capital per capita, share of foreign capital and share of state capital of an enterprise. For the second specification, in addition to those included in the first specification, it include city level variables could influence the establishment of APC, like age, tenure of mayor, GDP and share of secondary industry of the city and ect.

Table 1 give us descriptive statistic for variables included in specification 1 in year 2002. Descriptive statistic for additional variables include in specification 2 can be seen in appendix 7. As can be seen in Table 1, control group include 12799 firms and treat group include 23774 firms. And on average the response variable lnpatent00 is slightly higher in control group than treat group. Other control variables seems to be balanced between two group.

	control	treat
lnage	2.158	2.195
sd lnage	(0.911)	(0.977)
lnage2	5.488	5.771
sd lnage2	(4.158)	(4.511)
exp	0.307	0.229
sd exp	(0.461)	(0.420)
lnemp	4.883	4.910
sd lnemp	(1.141)	(1.143)
lnavek	4.748	4.674
sd lnavek	(1.227)	(1.137)
s_state	0.166	0.152
sd s_state	(0.354)	(0.341)
s_foreign	0.107	0.068
sd s_foreign	(0.29)	(0.23)
lnpatent00	2.587	2.250
sd lnpatent00	(20.438)	(18.795)
obs	12799	23774

Table 1: Descriptive statistic for treat and control group in year 2002 (specification 1)

Since the dataset used is panel data, it is intuitive to see how the response variable Y changes as time goes by. For each year, we can either calculate the average lnpatent per year or the aggregate lnpatent per year, which corresponding to panel left and panel right in Figure 1. Wang and Feng (2018) use aggregate lnpatent to show that sample satisfy **common trend** assumption, which i think is improper. The reason is that aggregate level would be influenced by the number of firms contained in each year. And the left panel in Figure 1 shows the average amount of lnpatent per year doesn't satisfy **common trend** assumption. What's more, before 2002 the average number of lnpatent is higher for control group, which lead to suspicion for the randomness of assignment assumption.

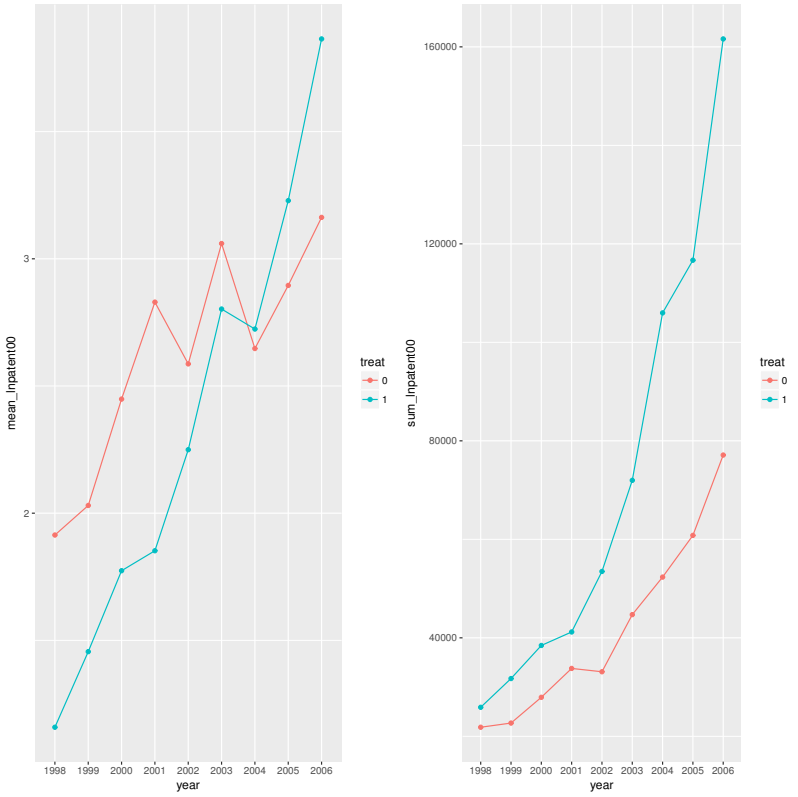

Figure 1: Left: Average lnpatents for both Treat group and Control group. Right; Aggregate lnatents for both Treat group and Control group

3 Double Machine Learning and Algorithm

As mentioned before, **common trend assumption** doesn't hold, which make the DID model not so sound. And linear specification for control variables may not adequately. Under such case, this paper use Double Machine Learning (DML) proposed by Chernozhukov et al. (2018). Considering the heterogeneous impact of APC on firms innovation, this paper adopt **Interactive Model** in DML. The model setting are as follows:

$$Y = g_0(D,X) + U, \qquad E_p[U|X,D] = 0 \tag{1}$$

$$D = m_0(X) + V, \qquad E_p(V|X) = 0 \tag{2}$$

And the parameter of interest in this model is the average treatment effect (ATE).

$$ATE : \theta_0 = E_p[g_0(1,X) - g_0(0,X)] \tag{3}$$

To estimate ATE, we need to set up moment condition.

$$E_p(\psi(W;\theta_0,\eta_0)) = 0 \tag{4}$$

$$\psi(W;\theta_0,\eta_0) = (g(1,X) - g(0,X)) + \frac{D(Y - g(1,X))}{m(X)} - \frac{(1-D)(Y - g(0,X))}{1 - m(X)} - \theta \tag{5}$$

DML use a semi-parametric residual-on-residual regression method for estimating average treatment effect and average treatment effect on treated. The basic idea behind DML estimator is to first run a non-parametric regression of outcomes on covariates using machine learning, and then run a second non-parametric regression of treatment assignment on covariates; Finally using the moment conditions with scores obeying orthogonality conditions to estimate parameter of interest.

The algorithm is shown as follows and the code for this paper can be seen in github(Using K=2):

Algorithm 1 Algorithm for Interactive Model in DML

1: Split Data into K =2 samples: main sample I^a and auxiliary sample I.
2: Train $Y_{i,D=1} = \hat{g_0}(D = 1, X_i) + \hat{U}_i$ using machine learning method, with $i_{D=1} \in I^a$
3: Train $Y_{i,D=0} = \hat{g_0}(D = 0, X_i) + \hat{U}_i$ using machine learning method , with $i_{D=0} \in I^a$
4: Train $D_i = \hat{m}_0(X_i) + \hat{U}_i$, with $i \in I^a$
5: Estimate $\hat{Y}(1, X_i) = \hat{g_0}(D = 1, X_i)$, with $i \in I$
6: Estimate $\hat{Y}(0, X_i) = \hat{g_0}(D = 0, X_i)$, with $i \in I$
7: Estimate $\hat{D}_i = \hat{m}_0(X_i)$, with $i \in I$
8: Estimate $\hat{\theta}_0(I, I^a) = \frac{1}{n_I}\sum_{i=1}^{n_I}(\hat{g}(1,X_i) - \hat{g}(0,X_i)) + \frac{D_i(Y_i - \hat{g}(1,X_i))}{\hat{m}(X_i)} - \frac{(1-D_i)(Y_i - \hat{g}(0,X_i))}{1-\hat{m}(X_i)}$
9: Switch sample I^a and I.
10: Repeat steps 2 to 6.
11: Estimate $\tilde{\theta}_0 = \frac{1}{2}(\hat{\theta}_0(I, I^a) + \hat{\theta}_0(I^a, I))$
12: Repeat steps 1 to 11 M times and average the resulting $\tilde{\theta}_0$.

4 Results

Using interactive model of DML described above, this paper consider 6 different machine learning methods. These machine learning methods include 2 tree-based methods , labeled 'Random Forest', 'Gradient Boost' and 4 regression-based methods, labeled as 'Lasso','Ridge','Elastic Net' and 'Neural Network'. As we mentioned before, the dataset is panel data, which include 387722 observations with 166324 treated and 226560 control. In additional to that, this paper also consider subset cross section samples (data in year 2002 ,2004). In data section, this paper introduce 2 specifications, which corresponding to two sets of control variables. For number of split , this paper consider K= 2 and K=5. To reduce the disproportionate impact of extreme propensity score weights in the interactive model, this paper trim the propensity score at 0.1 and 0.9. So different results are different setting combinations of the following :

1. K= 2, 5
2. Specification 1 and Specification 2 (Control variables included)
3. Full sample (panel data) and Subsample (data in year 2002)

And each result table report the following parameters:

1. Estimation of ATE and it's standard deviation Denoted as se(ATE)
2. The predictive mean square error of Y in validation set. Denoted as mse.v
3. The logloss of D in validation set. Denoted as logloss.v

Table 2 and Table 3 show results for panel data and cross section data under specification 1 and number of iteration M=1. Estimators with all machine learning methods except Neural Net have consistent results for both K=2 and K=5. However, either Table 2 or Table 3 shows that the results obtained from the different machine learning methods are consistent with each other. In table 2 , if we ignore result from Neural Net, ATE is not significant for Random Forest and is significantly positive also close to result in Wang and Feng (2018) (0.9330) for the remaining ML methods. Table 3 is for cross sectional data in year 2002, most of the estimators are not significant which is not surprising since figure 1 shows average Y (lnpatent00) are higher in control group than treat group. Finally, we

can find Neural Network perform worst in predicting out of sample treatment assignment D, which may relate to the unstable performance of ATE.

	Random Forest	Gradient Boost	Lasso	Ridge	Elastic Net	Neural Net
K=2						
ATE	-0.037	0.87	0.516	0.487	0.545	-0.715
se(ATE)	(0.102)	(0.121)	(0.078)	(0.078)	(0.078)	(0.13)
mse.v	(19.496)	(19.9)	(20.543)	(20.542)	(20.543)	(20.279)
logloss.v	(0.633)	(0.683)	(0.652)	(0.652)	(0.652)	(0.778)
K=5						
ATE	-0.066	0.882	0.542	0.481	0.562	1.859
se(ATE)	(0.1)	(0.123)	(0.078)	(0.078)	(0.078)	(0.095)
mse.v	(19.108)	(19.815)	(20.541)	(20.539)	(20.542)	(20.273)
logloss.v	(0.623)	(0.675)	(0.652)	(0.652)	(0.652)	(0.697)

Table 2: Panel data,M=1,Specification 1

	Random Forest	Gradient Boost	Lasso	Ridge	Elastic Net	Neural Net
K=2						
ATE	0.552	0.094	0.022	-0.106	-0.092	1.153
se(ATE)	(0.334)	(0.398)	(0.258)	(0.258)	(0.257)	(0.359)
mse.v	(19.19)	(19.052)	(19.007)	(18.993)	(18.994)	(18.886)
logloss.v	(0.654)	(0.717)	(0.636)	(0.636)	(0.636)	(0.693)
K=5						
ATE	0.523	0.126	0.04	-0.093	-0.065	0.305
se(ATE)	(0.346)	(0.411)	(0.258)	(0.259)	(0.257)	(0.405)
mse.v	(19.288)	(19.047)	(18.997)	(18.993)	(18.993)	(18.904)
logloss.v	(0.652)	(0.709)	(0.636)	(0.636)	(0.636)	(0.817)

Table 3: Cross section (Year = 2002), M=1, Specification 1

Results for specification 2 and also for cross section sample (year =2004) can be seen in APPENDIX.

References

Blind, K. (2012). The impact of regulation oninnovation. Nesta Working Paper Series. 2

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68. 2, 5

Wang, Y. j. and Feng, X. (2018). The reform of administration approval system and firms' innovation. *China's Industrial Economics*. 2, 3, 4, 7

Appendices

A Tables

	Random Forest	Gradient Boost	Lasso	Ridge	Elastic Net	Neural Net
K=2						
ATE	3.534	-0.386	0.097	0.052	0.228	0.477
se(ATE)	(0.706)	(0.075)	(0.078)	(0.078)	(0.078)	(0.176)
mse.v	(19.669)	(19.672)	(20.544)	(20.543)	(20.543)	(20.184)
logloss.v	(0.632)	(0.649)	(0.652)	(0.652)	(0.652)	(0.815)
K=5						
ATE	2.673	-0.411	0.101	0.065	0.248	2.531
se(ATE)	(1.368)	(0.074)	(0.078)	(0.078)	(0.078)	(0.119)
mse.v	(19.437)	(19.317)	(20.54)	(20.538)	(20.541)	(20.217)
logloss.v	(0.623)	(0.649)	(0.652)	(0.652)	(0.652)	(0.73)

Table 4: Pannel data , M=1, Specification 2

	Random Forest	Gradient Boost	Lasso	Ridge	Elastic Net	Neural Net
K=2						
ATE	-0.044	0.193	0.308	0.48	0.471	0.51
se(ATE)	(0.279)	(0.411)	(0.257)	(0.257)	(0.256)	(0.872)
mse.v	(19.303)	(19.195)	(18.996)	(18.991)	(18.992)	(18.98)
logloss.v	(0.631)	(0.719)	(0.636)	(0.636)	(0.636)	(0.989)
K=5						
ATE	-0.058	-0.235	0.281	0.409	0.39	-0.638
se(ATE)	(0.273)	(0.391)	(0.257)	(0.258)	(0.257)	(0.276)
mse.v	(19.013)	(19.02)	(19.002)	(18.993)	(18.996)	(18.955)
logloss.v	(0.628)	(0.713)	(0.636)	(0.636)	(0.636)	(0.825)

Table 5: Cross section (year =2002),M=1,Specification 2

	Random Forest	Gradient Boost	Lasso	Ridge	Elastic Net	Neural Net
K=2						
ATE	0.288	0.394	1.283	1.547	1.553	2.62
se(ATE)	(0.259)	(0.353)	(0.244)	(0.246)	(0.244)	(0.812)
mse.v	(21.02)	(21.078)	(20.999)	(21.001)	(21)	(20.964)
logloss.v	(0.615)	(0.699)	(0.623)	(0.623)	(0.623)	(0.793)
K=5						
ATE	0.367	-0.088	1.217	1.468	1.375	0.191
se(ATE)	(0.253)	(0.335)	(0.242)	(0.245)	(0.242)	(0.376)
mse.v	(20.853)	(21.013)	(20.982)	(20.98)	(20.978)	(20.924)
logloss.v	(0.614)	(0.69)	(0.623)	(0.623)	(0.623)	(0.657)

Table 6: Cross section (year=2004) M=1, Specification 2

	control	treat
age_ps_post	0.599	0.207
sd age_ps_post	(0.490)	(0.405)
tenure_mayor_post	1.098	1.712
sd tenure_mayor_post	(0.964)	(2.156)
neibor_post	0.112	0.273
sd neibor_post	(0.105)	(0.248)
age_ps_t	2.996	1.035
sd age_ps_t	(2.450)	(2.026)
age_ps_t2	14.978	5.177
sd age_ps_t2	(12.252)	(10.130)
age_ps_t3	74.889	25.884
sd age_ps_t3	(61.262)	(50.652)
tenure_mayor_t	5.491	8.561
sd tenure_mayor_t	(4.818)	(10.779)
tenure_mayor_t2	27.453	42.804
sd tenure_mayor_t2	(24.092)	(53.893)
tenure_mayor_t3	137.267	214.021
sd tenure_mayor_t3	(120.458)	(269.466)
neibor_t	0.559	1.366
sd neibor_t	(0.524)	(1.241)
neibor_t2	2.796	6.830
sd neibor_t2	(2.620)	(6.206)
neibor_t3	13.978	34.149
sd neibor_t3	(13.099)	(31.028)
Industrial_t	0.427	0.563
sd Industrial_t	(0.912)	(1.075)
Industrial_t2	2.136	2.816
sd Industrial_t2	(4.560)	(5.374)
Industrial_t3	10.680	14.078
sd Industrial_t3	(22.802)	(26.869)
lnfirm_t	32.679	33.319
sd lnfirm_t	(5.093)	(4.466)
lnfirm_t2	163.396	166.594
sd lnfirm_t2	(25.464)	(22.328)
lnfirm_t3	816.982	832.970
sd lnfirm_t3	(127.322)	(111.638)
obs	12799	23774

Table 7: data descriptive for additional control variables

B Code

```
getresult_commonsupport<-function(data,K,yvar,xvar,xvar_d,d,Methods,M){
  nmethod <- length(Methods)
  ###create empty tables for results
  ate.table <- matrix(0,1,nmethod)
  se.ate.table <- matrix(0,1,nmethod)
  mse.out.table <- matrix(0,1,nmethod)
  logloss.out.table <- matrix(0,1,nmethod)
  h2o.data <- as.h2o(data)
  ####for all machine learning methods in Methods
  for ( m in (1:M)){
    split <- runif(nrow(data))
    cvgroup<-as.numeric(cut(split,
    breaks = quantile(split,probs = seq(0,1,1/K)),
    include.lowest = T))

    ####for all machine learning methods in Methods
    for (i in (1:nmethod)){
      method <- Methods[i]

      ate<- 0
      se.ate <- 0
      mse.out <- 0
      logloss.out <- 0

      for ( j in 1:K){
        ii <-which(cvgroup == j)
        nii <-which (cvgroup != j )
        datause = h2o.data[nii,]
        dataout = h2o.data[ii,]
        calxx<- ML(datause = datause, dataout = dataout,yvar= yvar,d=d,
        method = method,xvar = xvar,xvar_d = xvar_d)
        yout = as.matrix(dataout[,yvar] )
        dout =as.matrix(as.numeric(dataout[,d]$ALC)-1)
```

```
        atedata <- cbind(yout,dout,calxx$my_d1x,calxx$my_d0x,calxx$my_x)

        ###common support : keep only observations with (0.1<p<0.9)
        atedata <- atedata[ which(atedata[,5]>0.1 &atedata[,5]<0.9), ]




        ate <- ate + ATE(atedata[,1],atedata[,2],atedata[,3],
        atedata[,4],atedata[,5])
        se.ate <-se.ate + (SE.ATE(atedata[,1],atedata[,2],
        atedata[,3],atedata[,4],atedata[,5]))^2
        mse.out <- mse.out+ calxx$pymse
        logloss.out <- logloss.out + calxx$logloss_d
      }

      ate.table[,i] <-ate.table[,i]+ ate/K
      se.ate.table[,i] <- se.ate.table[,i]+sqrt(se.ate/(K^2))
      mse.out.table[,i] <- mse.out.table[,i]+ sqrt(mse.out/K )
      logloss.out.table[,i] <-logloss.out.table[,i]+logloss.out/K

    }
  }
  ate.table <- ate.table/M
  se.ate.table <- se.ate.table/M
  mse.out.table <- mse.out.table/M
  logloss.out.table <- logloss.out.table/M

  return(list(ate.table = ate.table,se.ate.table = se.ate.table,
  mse.out.table = mse.out.table,logloss.out.table = logloss.out.table))
}
```