Brief Final Project Proposal

CS112 - Fall 2019

Md Mahmudunnobe, Jackie Trang, Claudia Gold

Paper introduction

This paper replicates an original paper which analyzed the association of the increase in reporting the crime against female with the female empowerment. This paper classifies the original data into different states and shows that the reporting crime is increasing for all states around 1995 irrespective of the implementation year of the state, which weakens the claim of the original paper. They showed that the classification of 'crime against women' changes in 1995 which increases the aggregated number of crimes after 1995 in all states. They run the same regression model with only the data after 1995 and found that the coefficient is not statistically significant at this time. They further analyze the control variables to predict the implementation year for different states. We will dig deeper into the relationship between Indian state-level policy implementation and crimes against women by replications and extensions.

Paper Link:

https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/25677/MYLSD5&version=1.0

Code Link (in R):

https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/25677/MA6GIE&version=1.0

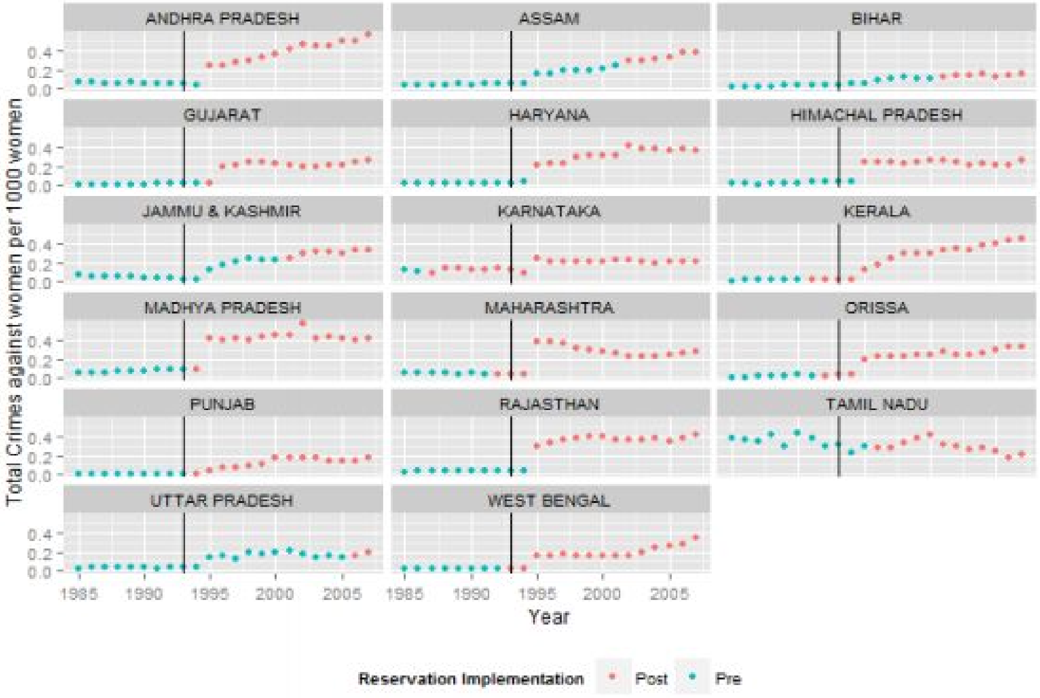
Data Link:

https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/25677/FEBCP8&version=1.0

Part I: Replication

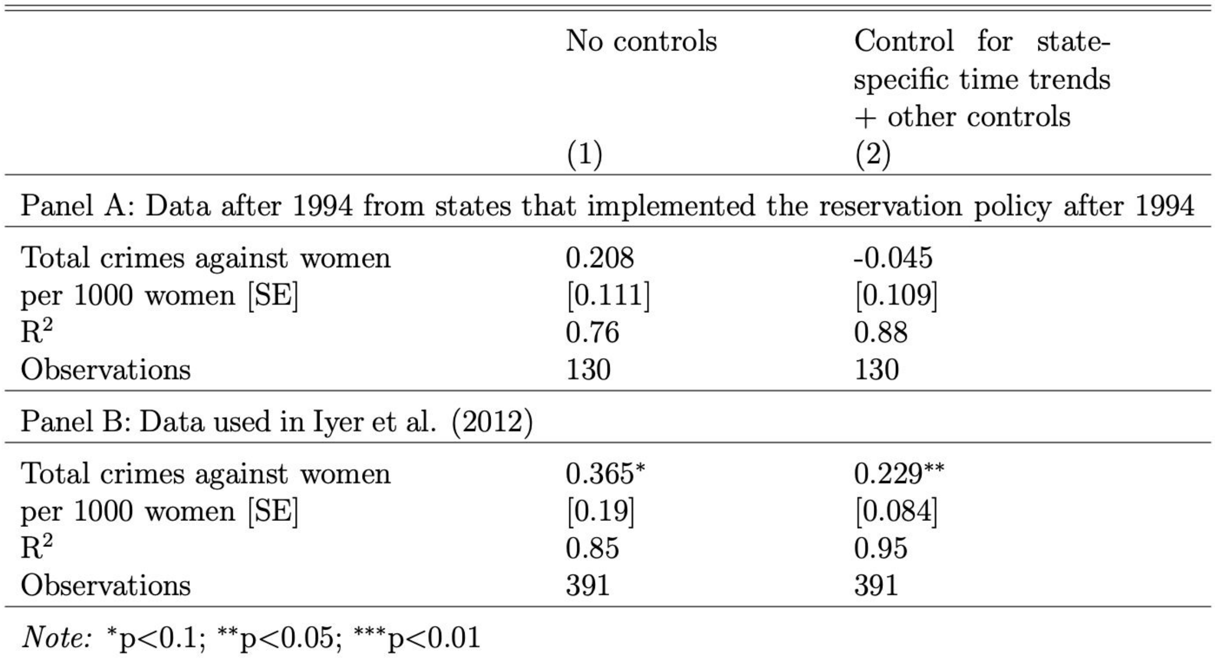
We will replicate the figure of total crimes against women per 1000 women across states in India, 1985-2007 and the table of women's political representation and crimes against women (Table 1). By replicating, we can show our understanding of the data, be clear about our purpose and the next move for the extensions.

Figure 2: Total crimes against women per 1000 women across states in India, 1985-2007



Note: Blue points indicate years prior to implementation of reservation rule at state-level, and red points indicate the first year of reservation and all years subsequent. Vertical line at 1993 denotes the year of ratification of the national amendment. Almost all states show a surprising spike in 1995, which does not usually correspond to either the national amendment ratification or the state-level implementation.

Table 1: Women's political representation and crimes against women



	No controls (1)	Control for state-specific time trends + other controls (2)
Panel A: Data after 1994 from states that implemented the reservation policy after 1994		
Total crimes against women	0.208	-0.045
per 1000 women [SE]	[0.111]	[0.109]
R^2	0.76	0.88
Observations	130	130
Panel B: Data used in Iyer et al. (2012)		
Total crimes against women	0.365*	0.229**
per 1000 women [SE]	[0.19]	[0.084]
R^2	0.85	0.95
Observations	391	391

Note: *p<0.1; **p<0.05; ***p<0.01

Part II: Extension

Extension 1: Run genetic matching to find the average treatment effect.

Data pre-processing: Divide the original dataset into two different sub-dataset with the year 1995 as a threshold. 1995 is the year that National Crime Records Bureau (NCRB) in India began classification of crimes into detailed categories of dowry death, molestation, sexual harassment, and cruelty by husband and/or his relatives (NCRB 1995). Therefore, the way crimes were classified before and after this year is different, which leads to bias if we treat them the same way.

- Subset 1 - before classification: Data before 1995
- Subset 2 - after classification: Data after 1995
- In each subset, we set the state-level policy implementation year of the quota on the proportion of women in legislative bodies as the threshold to indicate treatment status. This year might be

different from the national implementation year of this quota in 1992 because some states adopted this quota much sooner or much later than the national year.

- ○ Control group (treat = 0) is the data before the state implementation year.
- ○ Treatment group (treat = 1) is the data after the state implementation year.

Data analysis for Extension 1:

We will run Genetic Matching on each subset to arrive at the average treatment effect on the crime rate. The dependent variable is the reported crime rate against women and the independent variables are GDP per capita, Female-Male ratio, Overall Literacy, Women's Literacy and Percent of the rural population (these independent variables have been shown in table 2 below).

After we have the 2 average treatment effects (ATT) from genetic matching from the two subsets, we can get the average of this ATT to find a more holistic ATT for the whole dataset without the bias in inconsistent crime classification in 1995. However, if the two ATTs from the two subsets are completely different, we can arrive at meaningful insights. We will use the default settings for GenMatch as this paper didn't use Matching at all.

Extension 2: Run random forest

Random forest: We will use the available control variables (as shown in Table 2) in 1985 as our predictors and the year of the reservation policy implemented as our dependent variable or response to run a random forest. In the replication paper, they used these predictors in a regression model to understand how closely these parameters, which they believe as the indicators of progressivity of different states towards women, are associated with the earlier or later implementation of the policy. By running random forest, we want to see how small the OOB MSE is and which variables are more important than others to predict the implementation year.

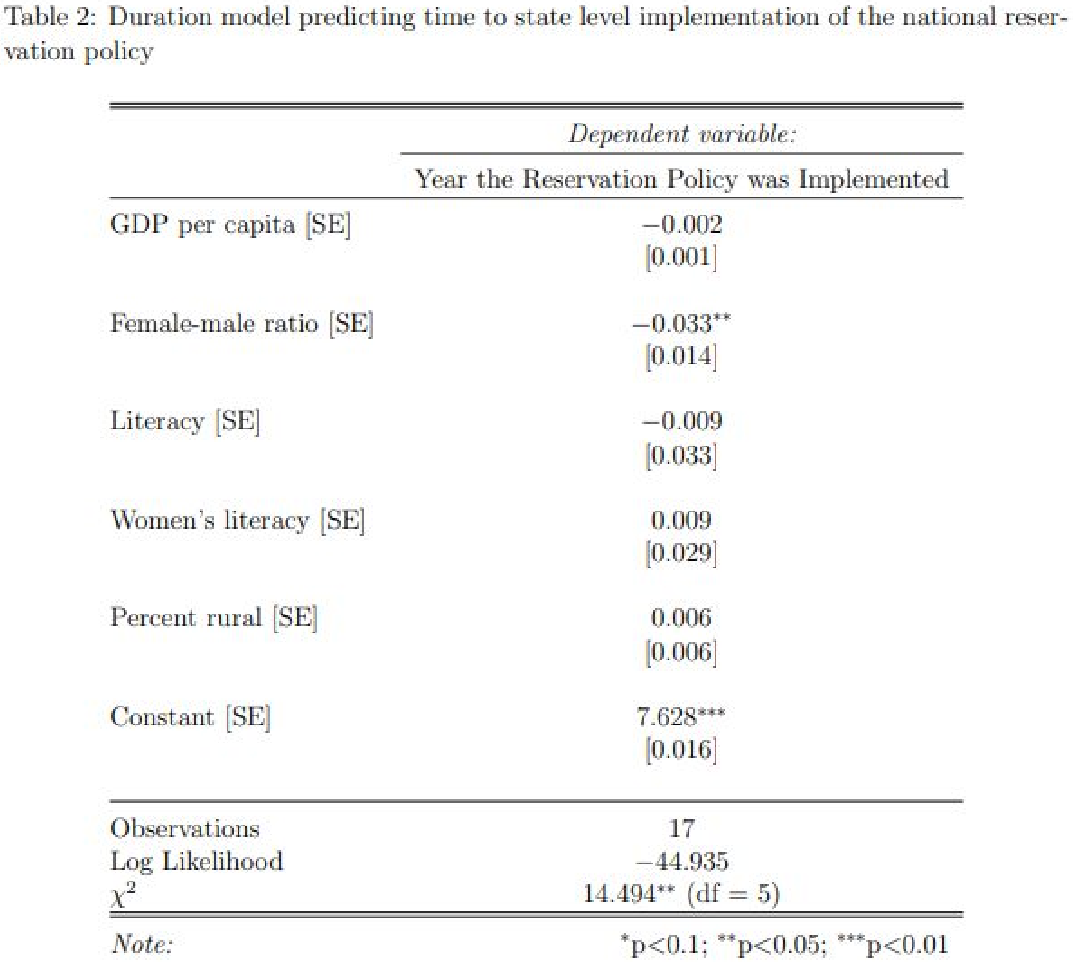


Table 2: Duration model predicting time to state level implementation of the national reservation policy

	Dependent variable:
	Year the Reservation Policy was Implemented
GDP per capita [SE]	−0.002 [0.001]
Female-male ratio [SE]	−0.033** [0.014]
Literacy [SE]	−0.009 [0.033]
Women's literacy [SE]	0.009 [0.029]
Percent rural [SE]	0.006 [0.006]
Constant [SE]	7.628*** [0.016]
Observations	17
Log Likelihood	−44.935
χ^2	14.494** (df = 5)
Note:	*p<0.1; **p<0.05; ***p<0.01

Advice seeking

We propose two options for extensions, but we are not sure which one is better to execute or we go with both. Column names of data are not easily understood but we are thinking to use the codes to understand the important variables. We are wondering if that will be harder or not.