Data mining on 'UIP' and immunization coverage in India

Dweepobotee Brahma[1]
Ph.D. student of Economics
Western Michigan University
Kalamazoo, MI 49008, USA
dweepobotee.brahma@wmich.edu

Debasri Mukherjee
Professor of Economics
Western Michigan University
Kalamazoo, MI 49008, USA
debasri.mukherjee@wmich.edu

Abstract

This paper examines the effect of Universal Immunization Program (UIP) on the coverage of essential vaccines in various Indian states. Our empirical findings confirm that UIP by itself does not always significantly affect the coverage, however, the effect of UIP on coverage becomes significantly positive in the presence of good health infrastructure. Health infrastructure not only helps the coverage by itself, but it also raises the effectiveness of vaccination funds on the coverage. Health infrastructure combined with UIP funds also serve as good predictors of immunization coverage based on application of LASSO and LARS variable selection techniques. We also compare the nonparametric kernel densities of response variable to that of its in-sample prediction and find that our model fits the data almost perfectly for each vaccine measures considered. The policy prescription that follows from our study is that the immunization program should focus on promoting the required infrastructure in addition to providing funds in order to facilitate effective usage of funds. Some states that are consistently underperforming require more targeted policies.

Keywords: Immunization in India, health infrastructure, variable selection and shrinkage, LASSO, Least Angle Regression, nonparametric density.

JEL classification: C14, C52, I15, O2

[1] Corresponding author email: dweepobotee.brahma@wmich.edu

1. Introduction:

India has a startlingly high incidence of infant and maternal mortality resulting from vaccine-preventable diseases, despite the high GDP growth rate. India has the highest number of births per year (about 26 million) in the world and alone contributes to about one-fifth of child mortality worldwide, mostly from such diseases.[2] Various social, cultural and economic factors compounded by supply-side factors (as identified by UNICEF, India) such as inefficient logistics, transportation facilities, and inadequate health care staff resulting in high levels of wastage of vaccines are responsible for the problem. The Universal Immunization Program (UIP) launched in 1985 is India's policy response to tackling this problem. UIP is also an essential component of India's efforts to meet the Millennium Development Goals, specifically goals 4 and 5, namely to reduce infant mortality and improve maternal health. However, as of 2015, the infant-mortality rate in India is still quite high, at about 38 per 1000 live births with a wide variation in the coverage of vaccinations as well as disbursement of funds across the different states in India.[3] Although several studies using household survey data have looked at the effect of household characteristics on immunization or the resulting anthropomorphic measures, to the best of our knowledge, empirical works investigating the impact of UIP expenditure on immunization coverage at a macro level is missing. Using Indian state-wide annual data (31 states and six years based on data availability) we examine the effect of UIP funds on immunization coverage of various vaccines and the role of health infrastructure in this context. We consider five vaccines (five outcome/dependent variables), namely BCG, polio (OPV), DPT, measles, TT (to be discussed in detail), and run standard regressions on various covariates for performing inference on them. We then employ two separate types of data mining techniques - (i) LASSO and LARS to check the predictive power of our covariates for each regression (ii) comparisons of nonparametric densities of observed and predicted values of the response variables to assess the overall fit of the modeling used. Our empirical data mining exercises robustly indicate that health infrastructure (to be defined in detail) is a crucial factor behind immunization and that its presence significantly facilitates the effectiveness of funds on coverage outcome. The upshot is that policy makers need to focus on health infrastructure as well, in addition to disbursement of more vaccination funds.

2. A brief background of immunization programs:

In 1978 Indian launched its Extended Program on Immunization (EPI) with the aim of immunizing all children with DPT, OPV, BCG and Typhoid by the first year of their lives. However, it was largely unsuccessful and managed to reach only urban areas. In 1985 India revamped and relaunched its immunization program-the Universal Immunization Program (UIP), this time in a phased manner aiming to cover all districts by 1989-90. This program established the cold chain logistic system for proper storage and transport of vaccines, and aimed to reduce mortality and morbidity from six vaccine-preventable diseases. It is one of the largest in the world with regard to the volume of vaccines used and number of recipients targeted. Currently, UIP is run by the

---

[2] See http://unicef.in/Whatwedo/3/Immunization

[3] While states like Karnataka and Andhra Pradesh report nearly 100% coverage under UIP, other states likes Madhya Pradesh and Rajasthan report only a 75% coverage. There is also a wide variation in the funds received by states under UIP. Chandigarh received about one lakh under UIP while its neighboring state Himachal Pradesh received more than 200 lakh in a year and the difference can't be explained by population size alone.

Immunization division under the National Rural Health Mission supervised by the Ministry of Health and Family Welfare. While the immunization division provides assistance to public health infrastructure units at various levels (state, districts, primary health centers (PHCs), community health centers (CHCs)) to undertake activities under UIP like routine immunization, training programs for health care staff and cold chain handlers, ensuring safe disposal practices, running awareness campaigns and monitoring adverse effects following immunization, UIP provides vaccines free of cost to all public health centers and private medical practitioners.[4] Although most health programs are funded and administered by state governments, UIP is funded entirely by the Central government and currently administers eight vaccines covering nine vaccine preventable diseases.

This study considers the five vaccines the program started with, namely BCG (Baccilus Calmette Guerin), OPV (Oral Polio Vaccine), Measles, DPT (diphtheria, tetanus and pertussis (whooping cough)) and TT (Tetanus Toxoid for pregnant women). BCG which protects against tuberculosis, requires a single dosage at birth administered intra-dermally. OPV and DPT both require five doses administered at specific ages up to 2 years and 5 years of age respectively. Measles requires two doses administered sub-cutaneous by 2 years of age. TT administered to pregnant women also requires two doses, with the first dose in the third trimester and the second dose four weeks after the first.

A similar, relatively younger policy under the National Rural Health Mission called Janani Suraksha Yojana (JSY) started in 2005 also aims at reducing maternal and infant mortality by integrating ante-natal care, institutional delivery and post-natal care. While JSY also includes TT vaccines to pregnant mothers and BCG vaccine to newborns, the cost of these vaccines is covered under UIP. Like the UIP, JSY is funded entirely by the Central Government. However, it operates through a financial incentive mechanism by providing conditional cash transfers to mothers at delivery and vouchers to community health workers. It is important to note here that while there exists some overlap between UIP and JSY in their goals and tools, they are two separate centrally funded schemes. The cost of administering vaccines (even within the JSY) is borne by UIP. This allows us to study the impact of UIP on immunization coverage without any contamination.

Since its establishment, UIP has been instrumental in reducing the incidence of infant and maternal mortality from vaccine-preventable diseases. Perhaps the most significant milestone has been the removal of India from WHO's watch list of countries where polio is considered an epidemic. Despite this, UIP is yet to achieve a 100% coverage all over India.

[4] For a more detailed description of the policy see the description by the Ministry of Health and Family Welfare at https://www.nhp.gov.in/sites/default/files/pdf/immunization_uip.pdf and also at http://www.itsu.org.in/about-UIP-in-india

3. Existing Literature.

The existing literature on UIP in the fields of economics, public policy and health policy concentrate on the short-term and long-term health effects on immunization exploiting survey data at the household or individual level. Patra (2006) uses household data from the National Family Health Survey-2 (1998-99) and estimates the effect of various demographic and socio-economic factors like income, education level, religion along with if the mother received any ante-natal care (binary variable) on the likelihood of child's immunization using a logistic regression. He finds increases in mother's education level, as well as exposure to pre-natal care to be associated with an increase in the likelihood of immunization. He classifies the different states in India into three categories, namely the 'Empowered Action Group', North-eastern states, and others.

Datar, Mukherji and Sood (2005) use the same dataset to estimate the effects of a constructed measure of health infrastructure (nearest primary health infrastructure facility available to a family and the presence of community health workers relevant to immunization) and community outreach programs on the vaccination status of a child. They found only a modest effect of the availability of health infrastructure on immunization coverage, although the size and capacity of the health infrastructure unit showed bigger effects on immunization coverage. Following them we include the total number of operating health centers in each state as a control covariate for health infrastructure in our regression. Using district level survey data, Anekwe and Kumar (2012) estimate the effect of being exposed to UIP (using a categorical variable) during the first year of a child's life on anthropomorphic outcomes and vaccination status. They find a positive effect of UIP on anthropomorphic measure but not on vaccination status in their sample. Carvalho et al (2014) have analyzed the effect of the other policy- Janani Suraksha Yojana on childhood immunization and other health indicators using District Level Household Survey data, and find a positive effect of conditional cash transfer on the proportion of fully vaccinated children at the district level, along with a positive impact on other post-natal health indicators. Banerjee et al (2010) conduct a clustered randomized controlled trial in 134 villages in rural Rajasthan and find that in a low-immunization coverage setting, improving the reliability and supply of services has only a modest effect on improving coverage. Small financial incentives combined with improved reliability had large positive impacts on immunization and was more cost effective. Thus, the existing literature investigates effects of household characteristics on immunization, or the impacts of immunizations on anthropomorphic measures of individual children using survey data on India. We do not find any work investigating the problem at cross-state macro level for India over a period of time and our work attempts to fill that gap.

4. Data and variables:

We use state-wise annual longitudinal data on coverage of five vaccines (BCG, DPT, Measles, OPV and TT) and total funds disbursed under UIP for the years 2005-2007 and 2009-2011 from Indiastat.com.[5] The Union territories of Andaman and Nicobar Islands, Lakshadweep, Dadra and Nagar Haveli, and Daman and Diu were dropped due to unavailability of data. Information on funds for the year 2008 was unavailable for all the states. A few other observations where the value of funds disbursed under UIP were missing were dropped from the analysis. The main regression analysis was conducted using 175 observations of 31 states, pooling over the years.

We run five different regressions, one for each vaccine where our response variable is the coverage (number of children vaccinated) in a state in a given year. For TT, the coverage refers to the total number of pregnant women instead. The data on state-wise release of funds under the Immunization Program (measured in rupees), is adjusted for inflation using the price level in 2011. We also control for the income level for each state by including the state-wise per capita net state domestic product at factor cost at constant prices. A key covariate in this study is our measure of health infrastructure. It is the total number of Primary Health Centers (PHCs) and Community Health Centers (CHCs) in a state in a given year since both PHCs and CHCs are responsible for administering vaccines under UIP. Although the total number of health centers vary substantially across states, is a very slow-moving variable (small or no changes) over the time period in our study. Another important covariate is the size of the population of the state. While there is a wide variation of population across the states, state-level data for this variable is not collected in a yearly basis. The most relevant and closest measure that we find for our purpose is the number of children under 6 years which was available only for the census year, 2011. We use the values for 2011 as repeated observations for each state to control for their population.[6]

The north-eastern states of India have substantial differences from the rest of the Indian states in terms of smallness in size, socio-economic performance, dominance of tribal population in many areas, and their geographical location (mostly separated from main land India with Bangladesh sited in between, and being positioned in the backdrops of Himalayan mountain roads and harsher climate, making it more difficult for funds/vaccines/personnel to reach in a timely manner). To account for this we include a north-east (NE) dummy in our specification which takes the value of 1 for the states of Assam, Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura, and zero for all other states. We also control for the states that have historically underperformed in terms of socio-economic indicators, and hence have been given a widely known special status – the so called "BIMARU" states (Bihar, Madhya Pradesh, Rajasthan and Uttar Pradesh). The dummy 'BIMARU' takes the value of 1 for these states and zero for all other states.

---

[5] We do not find data on more recent years, or specific funds disbursed for various vaccines.

[6] For TT vaccine targeted to pregnant women we use both (i) total female population as well as (ii) children under 6 years for relevant 'population' measure, and obtain qualitatively similar results. More appropriate measures are not available.

5. Estimation, variable selections and Results

We run several regressions for each of our five response (immunization) variables and our results are reported in Tables 3-7. Note that the first four vaccinations are applied to small children whereas the last one applies to pregnant mothers. The primary regression equation is:

$$Y = \beta X + U$$

where Y refers to dependent variable (coverage of a vaccine). X is the matrix of covariates, and U is the idiosyncratic error term. Our main covariates are UIP funds and health infrastructure and we add square terms for the funds to check for any possible diminishing returns. We also use interaction terms between fund and infrastructure to check for complementarity in how they affect immunization coverage. Our link function is linear with nonlinear covariates. After controlling for other covariates such as population of relevant age group, per capita state domestic product, regional dummies and interaction terms of the dummies with funds, we find that fund itself (or funds square also) may not always be statistically significant and we do not find any significant evidence of diminishing returns. However, infrastructure and its interaction term with fund always produce positive (as expected) and statistically highly significant coefficients. Infrastructure does always facilitate the immunization. The coefficient of the interaction term (infrastructure with funds) turns out to be (positive) statistically significant at less than 5 % level in all cases, indicating strong evidence of complementarity. Better infrastructure improves effectiveness of funds.

Time trend coefficient always turns negative and significant which is also reflected in Figures 1-3, indicating that over time number of people vaccinated has gone down (although we find number of children belonging to the age group has gone up and inflation adjusted funds have also gone up), a puzzle that needs some scrutiny. One possible explanation is the addition of Hepatitis-B vaccine in the UIP schedule not accompanied by a commensurate increase in the funds. The NE and Bimaru states do not always have significantly different immunization coverage from the rest of the areas. The coefficients of SDP and population variables are positive and significant as expected. [7]

Since there are many potential correlated variables in our study, variable selection is a matter of concern as in any macro-development data of similar types. We employ LASSO (least absolute shrinkage and selection operator) to examine covariates that can be selected as best predictors. LASSO achieves that by solving the following constrained optimization exercise.

---

[7] One may be concerned with the issue of heterogeneity. However, one of our most important covariates, i.e., health infrastructure, though varies substantially across states, is a slow-moving variable over the sample time periods that we study. Another important control covariate, child population variable is only available in 2011 census, hence a time invariant covariate. Since fixed effects model tends to drop such time invariant covariates due to collinearity and these are our important covariates to make inference on, we choose not to use fixed effects. The Wald tests for random effects give no conclusive result. Therefore, we resort to simple linear models, correcting for heteroscedasticity and adding some regional dummies. Our NE and BIMARU dummies used for the states that are most different from other states as a group, along with the aforementioned “slow-moving/time-invariant” covariates are likely to subdue any cross-state heterogeneities considerably. Our claim is corroborated by high R^2 produced in each regression, as well as kernel density plots depicting impressive model fitting.

$$\min\{(Y - X\beta)'(Y - X\beta)\} \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

where Y is the vector of outcome variable, X is the covariate vector, β is the coefficient parameter vector, $j=1,\ldots,p$ is the number of covariates, and s is a pre-specified tuning parameter that determines the amount of regularization. LASSO not only shrinks the coefficients but also *selects* the best predictors which is important for our purpose. Details of these methods can be found in Tibshirani et. al. (2008). The results from employing LASSO to the training set and its subsequent inference on the test set are shown in Table 9.

We also use LARS (least angle regression) technique for variable selection (and use Mallow's CP criteria for selection). The details are in Efron et al (2004). The basic steps for LARS are the following. To begin with, all the coefficients are set to zero. We start with the most correlated predictor of Y and increase its coefficient in the direction of its correlation with Y. We take the residual and continue increasing the coefficient until some other predictor has as much correlation with the residual. Then we increase both the coefficients in their joint least squares direction, until we find a third predictor with as much correlation with the residual. We continue this process until all the predictor are included in the model. It is argued in Efron et al (2004) that with minor modification and if standard regularity conditions are met, LASSO and LARS produce the same solution path as far as the selection and sparsity are concerned[8]. For each of our response variable both techniques agree on the set of relevant predictors. Our purpose behind employing these two techniques is to examine if the two variables funds and infrastructure get selected as good predictors of immunization. Both methods consistently select infrastructure variable, as well as interaction variable between fund and infrastructure as good predictors in every regression, along with population, as expected. Selections of other covariates vary among the immunization variables considered. Thus, our inference based on simple least square estimations - that infrastructure and its interaction with funds being highly significant - is also in compliance with LASSO and LARS selecting them as predictors for immunization coverage response.

The robustness of our conclusion above is also confirmed by data mining based on sample splitting method. We split our sample into two parts, treating first three years as training set and using the last three years as test sets. We use both LASSO and LARS techniques first to select relevant predictors from a superset of all covariates (as in Table 3 to 7) using 'training data' set on all states for first three years only, and then use that selected subset as covariates in the 'test data' set which consists of data on all states for the last three years. The new inference results are reported in Table 8 which corroborate the previous conclusion regarding the importance of infrastructure and its interaction with funds. Note that post selection inference on the same data set is not reliable. However, one can use selected predictors (based on training set) as covariates in the test set. Ideally one could also use three sets, training set, validation set and test set for such exercise. However, given our data limitation, we use only training set to "learn" about our predictors and then use them in the test set for further analysis (Table 8). Simple statistical "Machine learning" based on sample splitting acts as a good robustness check to our conclusion regarding the importance of our policy variables.

[8] In our study LARS uses Mallow's C_p to choose the covariates and LASSO uses %Deviance to choose the covariates. But both these methods select the same three covariates of Infrastructure, Infrastructure*Funds and the population.

To further investigate the heterogeneity across the states as well as their interaction with funds and infrastructure we perform variable selection using LASSO from a larger set of covariates consisting of dummies for each state[9], the interaction of these dummies with the funds allocated, the interaction of these dummies with the infrastructure variable, the interaction of these dummies with the funds allocated *and* infrastructure, as well as the other covariates from the previous specifications. With a total of 126 covariates, this gives us the most appropriate situation for employing LASSO where the number of covariates is much larger than the number of observations. We perform three-fold cross validation with the folds defined on each year, on our training data set to exploit the natural time-dependence structure of our data to select the tuning parameter. Using this tuning parameter, we perform variable selection on the training data set and subsequently inference on the test data set using the selected predictors. The regressions using these selected predictors show a good model fit as demonstrated by the R^2. The qualitative results from these regressions are presented in Table 10.

The most striking result is that the infrastructure (not funds allocated) and the target population are selected as predictors across all vaccines. By and large, the dummies for BIMARU states are selected along with their interactions with funds and infrastructure. The interaction of Bihar with funds shows a negative and statistically significant relation with coverage across all vaccines. This is likely due to the prevalence of certain religious dogmas that deter people from vaccinating their children.

Additionally, we try to assess the overall fit of the model from a different angle. Note that even in our not so large sample, R^2 consistently shows over 95% fit in every regression, attesting the choice of variables and modeling strategy used. To check how distribution of our in-sample prediction variable from each regression compares to that of the associated response variable, we plot nonparametric kernel densities of the two and report them in Figures 5-9.[10] The predicted variables in these plots are based on the results reported in the last columns of Tables 3 to 7. The plots show that the distributions match very well in all ranges of data.

6. Concluding Remarks:

We try to assess the importance of UIP fund for vaccination and heath infrastructure, the two policy variables that can be used for attaining full immunization coverage for children and pregnant mothers in India, a country still swamped with high infant mortality rate from vaccine preventable diseases. It is well known that inference and prediction are two (often contrasting) econometric goals with typical bias-variance trade-off (estimation strategies focusing on inference and hence targeted at bias reduction may well be associated with less parsimony and predictive power, and vice-versa). Our least square regressions show high statistical significance of the two policy variables. In a separate analysis, LASSO and LARS choose them as good predictors as well. When we split the sample into two sets, these two policy variables also get chosen by LASSO and LARS as predictors. We further obtain statistically significant coefficient for the policy variables in our

[9] After careful consideration we set Punjab as a reference state since it was very close to the median performance of all the states.

[10] We consider cross-validated bandwidths and Gaussian kernel function for nonparametric density plots.

test data as well. Overall nonparametric densities show excellent model fitting too. Therefore, infrastructure and its interaction with funds can be targeted as important policy variables for achieving our goals and the conclusion is supported from both inference as well as prediction based points of views. From the second set of LASSO results we find that infrastructure is selected as a predictor for coverage of the vaccines. The selection of ‘BIMARU’ state dummies and their various interaction terms reveal that these states are still underperforming after so many years of implementing UIP. This calls for special attention to these underperforming states from policy makers.

It is important to remember here that our infrastructure variable measured the total number of health centers in each state. The maintenance and staffing of these health centers are not under the purview of UIP. However, our results indicate that increasing funds under UIP will not be fruitful if there aren’t enough health centers to administer these vaccines. Better infrastructure not only facilitates immunization, it also makes funds more effective for the same purpose.

References

1. Banerjee Abhijit Vinayak, Duflo Esther, Glennerster Rachel, Kothari Dhruva. (2010) " Improving immunization coverage in rural India: clustered randomized controlled evaluation of immunization campaigns with and without incentives" *BMJ* 2010; 340 :c2220
2. Carvalho N, Thacker N, Gupta SS, Salomon JA (2014) More Evidence on the Impact of India's Conditional Cash Transfer Program, Janani Suraksha Yojana: Quasi-Experimental Evaluation of the Effects on Childhood Immunization and Other Reproductive and Child Health Outcomes. PLoS ONE 9(10): e109311. doi:10.1371/journal.pone.0109311
3. Datar, Ashlesha, Arnab Mukherji and Neeraj Sood. Health Infrastructure and Immunization Coverage in Rural India. Santa Monica, CA: RAND Corporation, 2005. http://www.rand.org/pubs/working_papers/WR294.html.
4. Efron, Bradley, Trevor Hastie, Iain Johnstone and Robert Tibshirani. "Least Angle Regression" *The Annals of Statistics* (2004), 32(2), 407-499.
5. Hastie, Trevor, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: data mining, inference and prediction* (2008). Springer.
6. Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, N.J: Princeton University Press.
7. Patra, Nilanjan, Universal Immunization Programme in India: The Determinants of Childhood Immunization. Available at SSRN: https://ssrn.com/abstract=881224
8. Tobenna D. Anekwe, Santosh Kumar; The effect of a vaccination program on child anthropometry: evidence from India's Universal Immunization Program. *J Public Health* 2012; 34 (4): 489-497. doi: 10.1093/pubmed/fds032

APPENDIX

Table1: Descriptive statistics of all the variables

	Minimum	Maximum	Mean	Median	Standard Deviation
B.C.G	8045	5693227	834455.4	539523	1094819
Measles	8590	5330048	763465.6	510755	1002537
O.P.V	8105	5544695	781922.5	524453	1032407
T.T	7260	5689521	790056.1	519934	1048690
Real Funds	63822	353538679	40562010	15583194	58903317
Per-capita SDP	7588	129397	39687	34096	22730
Infrastructure	1	4207	893	548	940.8
Under 6 pop	64111	30791331	5401406	3380721	6720341
Female pop	287507	95331831	19182168	12712303	21751093

The target group for TT vaccine is pregnant women. The target group for all other vaccines are children.

Table 2: Correlation between the vaccines considered in the dataset.

	B.C.G	**DPT**	**Measles**	**O.P.V**	**T.T.**
B.C.G	**1.00**				
DPT	0.99	**1.00**			
Measles	0.99	**0.99**	**1.00**		
O.P.V.	0.83	0.99	0.84	**1.00**	
T.T.	0.99	0.99	0.99	0.83	**1.00**

Figure 1:National trends in vaccine coverage

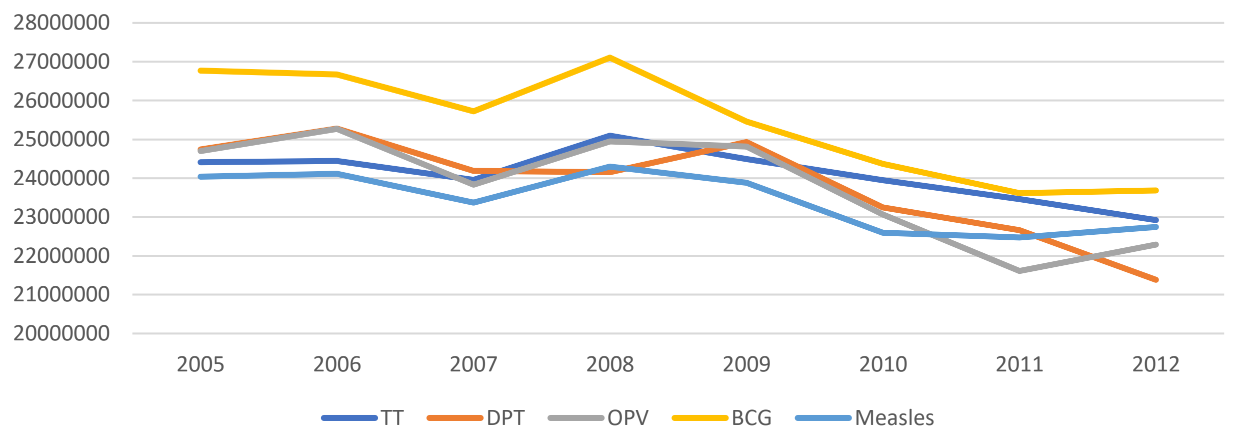


Figure 2:Vaccine trends for North-Eastern states

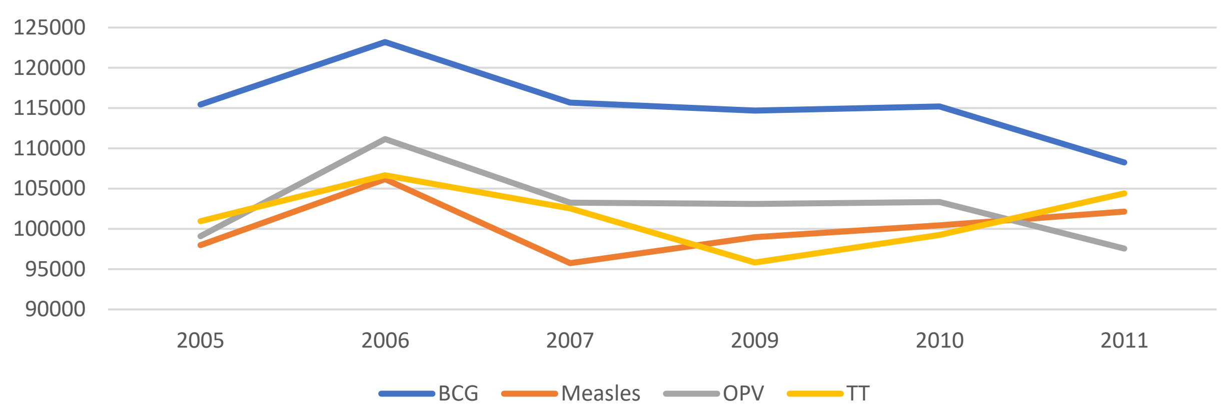


Figure 3:Vaccine trends for the BIMARU states

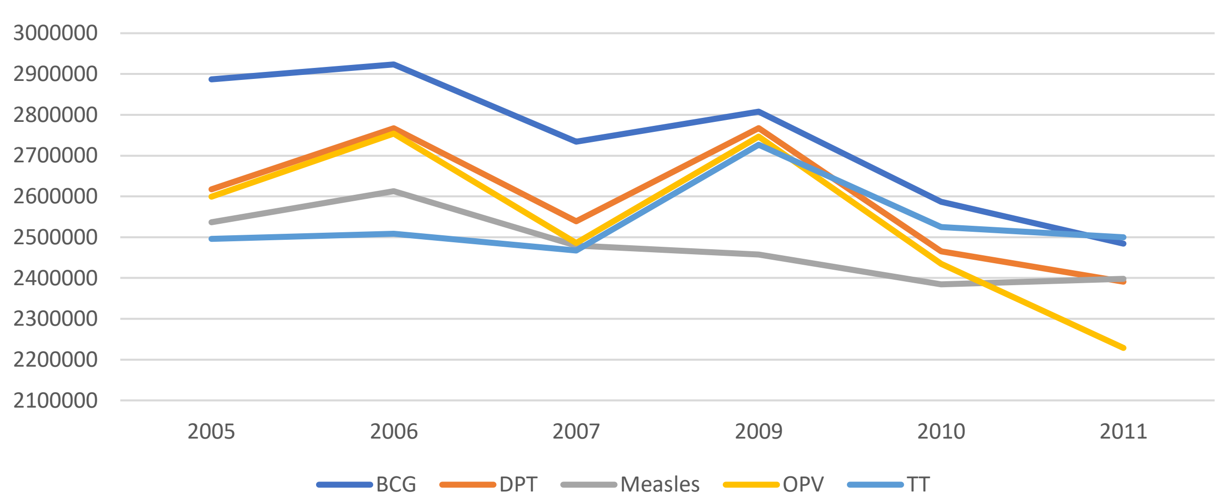


Table 3: Dependent Variable is BCG coverage

	(1)	(2)	(3)
Funds	-.00081 (0.397)	-.00086 (0.368)	-.00133 (0.306)
Infrastructure	144.13** (0.028)	133.66** (0.039)	109.52* (0.068)
SDP	.9434*** (0.006)	.7921* (0.063)	.5501 (0.358)
Population	.1197*** (0.000)	.1238*** (0.000)	.1239*** (0.000)
Funds*Infrastructure	2.01e-06*** (0.007)	1.98e-06*** (0.008)	2.31e-06** (0.020)
Trend	-27108.29*** (0.002)	-26256.07*** (0.003)	-25879.71*** (0.006)
$Funds^2$	-1.33e-11 (0.125)	-1.28e-11 (0.140)	-9.89e-12 (0.328)
NE	__	129.86 (0.995)	-23327.14 (0.453)
Bimaru	__	-63975.08 (0.458)	92171.32 (0.594)
NE*Funds	__	__	.00061 (0.378)
Bimaru*Funds	__	__	-.00204 (0.374)
constant	60542.89 (0.020)	62013.95 (0.081)	86764.51 (0.067)
R^2	0.9734	0.9735	0.9743

The numbers in parenthesis refers to p-values
*, **, *** indicates 10%, 5%,1% significance level respectively
Heteroscedasticity corrected robust standard errors have been used.

Table 4: Dependent Variable is DPT coverage

	(1)	(2)	(3)
Funds	-.0009381 (0.294)	-.0009405 (0.305)	-.0013444 (0.266)
Infrastructure	223.5849*** (0.004)	224.315*** (0.003)	200.427 *** (0.005)
Funds*Infrastructure	1.89e-06** (0.016)	1.88e-06** (0.015)	2.20e-06** (0.039)
$Funds^2$	-1.12e-11 (0.225)	-1.11e-11 (0.223)	-8.06e-12 (0.422)
SDP	1.078689*** (0.001)	1.112058** (0.014)	.8727198 (0.172)
Population	.1012944*** (0.000)	.1017015*** (0.000)	.1017233*** (0.000)
Trend	-22340.29** (0.012)	-22414.89** (0.017)	-22191.01** (0.020)
NE	—	4530.976 (0.842)	-13831.98 (0.680)
Bimaru	—	-4134.733 (0.960)	156447.8 (0.337)
NE*Funds	—	—	.0003501 (0.632)
Bimaru*Funds	—	—	-.0021289 (0.353)
Constant	28932.74 (0.298)	25023.24 (0.500)	48193.79 (0.353)
R^2	0.9654	0.9654	0.9663

The numbers in parenthesis refers to p-values.
*, **, *** indicates 10%, 5%,1% significance level respectively
Heteroscedasticity corrected robust standard errors have been used

Table 5: Dependent Variable is OPV coverage

	(1)	(2)	(3)
Funds	-.0011 (0.254)	-.0011 (0.266)	-.0016 (0.218)
Infrastructure	261.86*** (0.001)	263.64*** (0.001)	236.63*** (0.001)
SDP	1.318*** (0.001)	1.375*** (0.007)	1.105*** (0.103)
Population	.0913*** (0.000)	.09157*** (0.000)	.0916*** (0.000)
Funds*Infrastructure	2.18e-06*** (0.006)	2.16e-06*** (0.005)	2.53e-06** (0.018)
Trend	-29198.78*** (0.006)	-29364.12*** (0.008)	-29096.22*** (0.010)
$Funds^2$	-1.30e-11 (0.169)	-1.28e-11 (0.168)	-9.43e-12 (0.365)
NE	__	6043.44 (0.805)	-15204.7 (0.661)
Bimaru	__	-594.73 (0.995)	180431.5 (0.282)
NE*Funds	__	__	.0004206 (0.566)
Bimaru*Funds	__	__	-.00239 (0.295)
constant	45619 (0.149)	40282.52 (0.327)	66619.87 (0.225)
R^2	0.9605	0.9605	0.9616

The numbers in parenthesis refers to p-values
*, **, *** indicates 10%, 5%, 1% significance level respectively
Heteroscedasticity corrected robust standard errors have been used.

Table 6: Dependent variable is Measles coverage

	(1)	(2)	(3)
Funds	.0003338 (0.814)	.0003449 (0.809)	.000289 (0.875)
Infrastructure	228.84*** (0.004)	226.6958*** (0.004)	222.199*** (0.005)
SDP	1.244** (0.004)	1.1314** (0.031)	1.0864 (0.122)
Population	.08617*** (0.000)	.084533*** (0.000)	.0845*** (0.000)
Funds*Infrastructure	2.20e-06** (0.014)	2.25e-06** (0.012)	2.31e-06** (0.023)
Trend	-27241.37** (0.018)	-27012.09** (0.021)	-27017.35** (0.025)
$Funds^2$	-1.65e-11 (0.138)	-1.69e-11 (0.128)	-1.63e-11 (0.203)
NE	—	-16211.92 (0.542)	-18122.25 (0.584)
Bimaru	—	17609.74 (0.853)	49786.83 (0.823)
NE*Funds	—	—	-.0000138 (0.985)
Bimaru*Funds	—	—	-.0004341 (0.876)
Constant	39189.4 (0.243)	53107.85 (0.273)	57049.48 (0.287)
R^2	0.9564	0.9564	0.9565

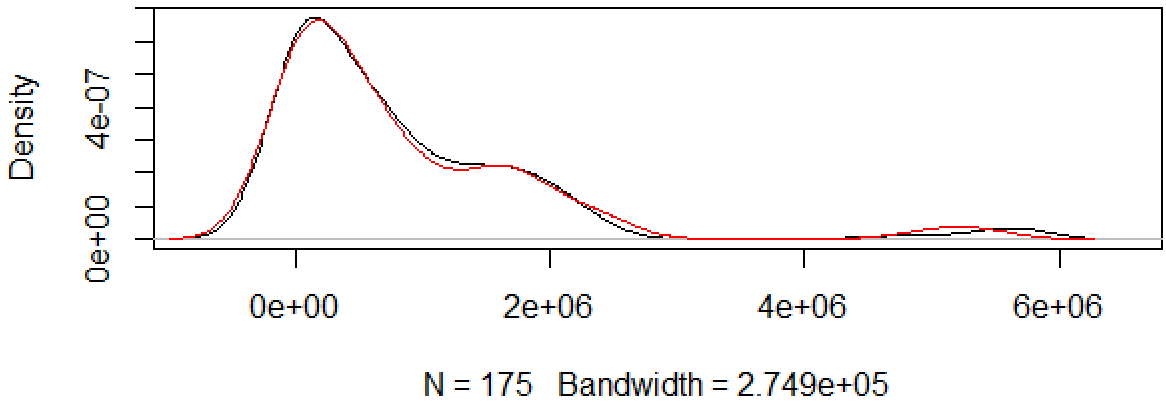
The numbers in parenthesis refers to p-values.
*, **, *** indicates 10%, 5%, 1% significance level respectively
Heteroscedasticity corrected robust standard errors have been used

Table 7: Dependent Variable is TT coverage

	(1)	(2)	(3)
Funds	-.0013 (0.245)	-.00122 (0.278)	-.00153 (0.303)
Infrastructure	417.44*** (0.000)	435.48*** (0.000)	411.05*** (0.000)
SDP	1.951*** (0.000)	2.338*** (0.000)	2.092** (0.010)
Population	.0686*** (0.001)	.0653*** (0.001)	.0652*** (0.002)
Funds*Infrastructure	2.05e-06** (0.011)	2.02e-06** (0.010)	2.34e-06** (0.030)
Trend	-23631.11** (0.027)	-25183.75** (0.027)	-25181.71** (0.032)
$Funds^2$	-8.06e-12 (0.377)	-8.09e-12 (0.364)	-4.73e-12 (0.642)
NE	__	24702.49 (0.381)	13322.77 (0.729)
Bimaru	__	63103.96 (0.575)	236715.2 (0.262)
Ne*Funds	__	__	-.000023 (0.977)
Bimaru*Funds	__	__	-.002337 (0.371)
constant	-14355.92 (0.695)	-37793.47 (0.449)	-16101.32 (0.794)
R^2	0.9464	0.9466	0.9475

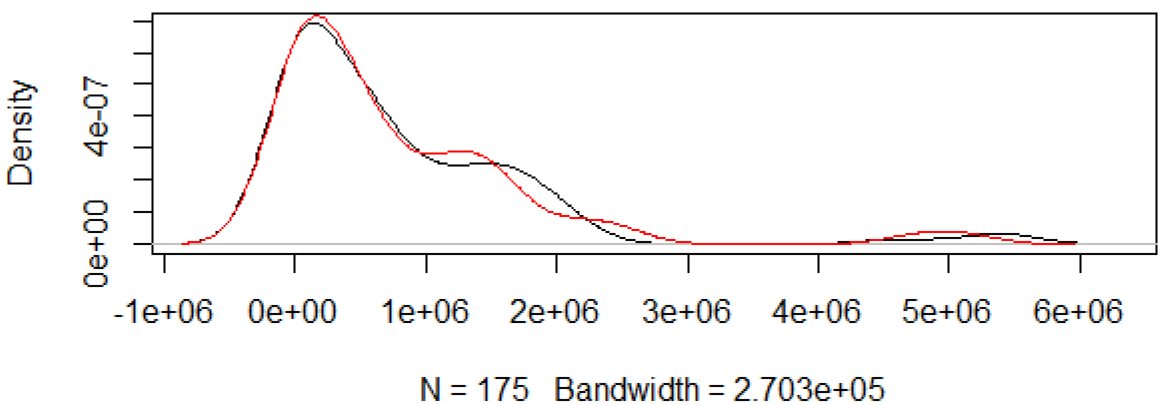
The numbers in parenthesis refers to p-values.
*, **, *** indicates 10%, 5%, 1% significance level respectively
Heteroscedasticity corrected robust standard errors have been used

Figure 5: Kernel density comparison for BCG



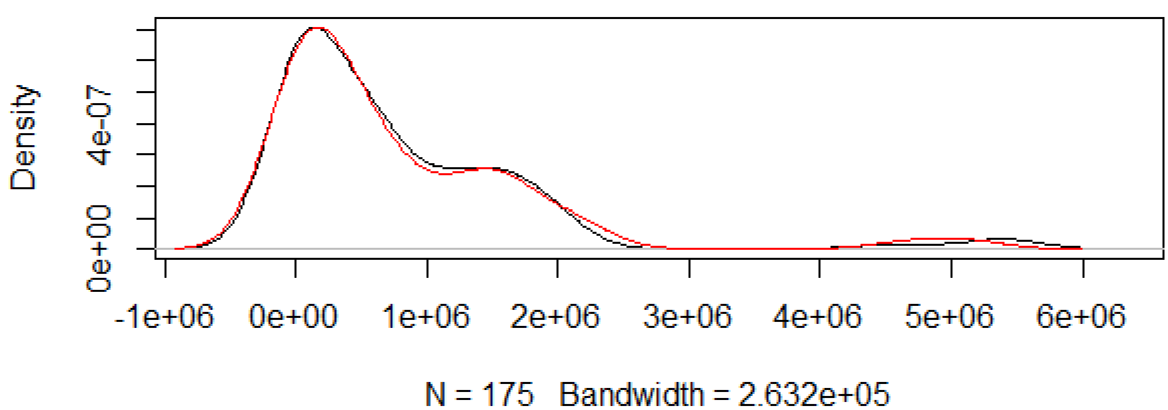
The black curve refers to the observed values of the dependent variable and the red curve refers to the predicted values of the same based on column (3) of Table 3.

Figure 6: Kernel Density comparison for DPT



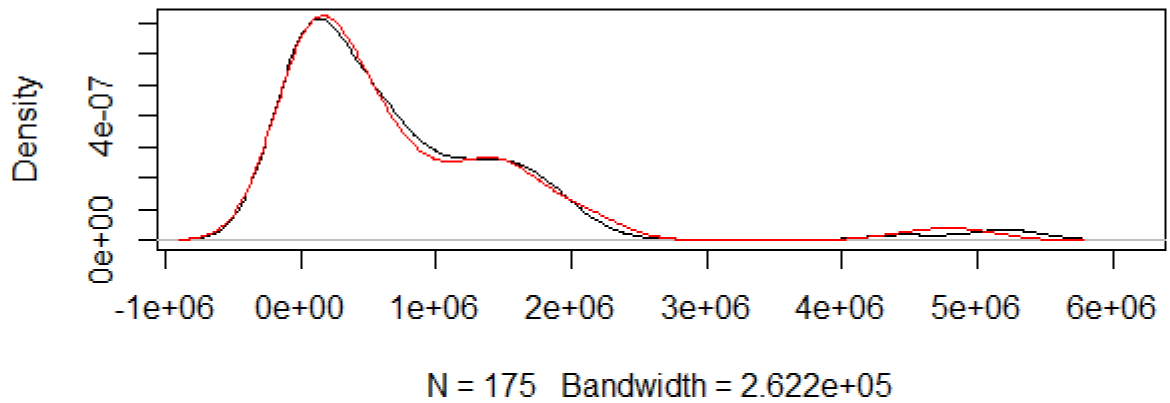
The black curve refers to the observed values of the dependent variable and the red curve refers to the predicted values of the same based on column (3) of Table 4.

Figure 7: Kernel density comparison for OPV



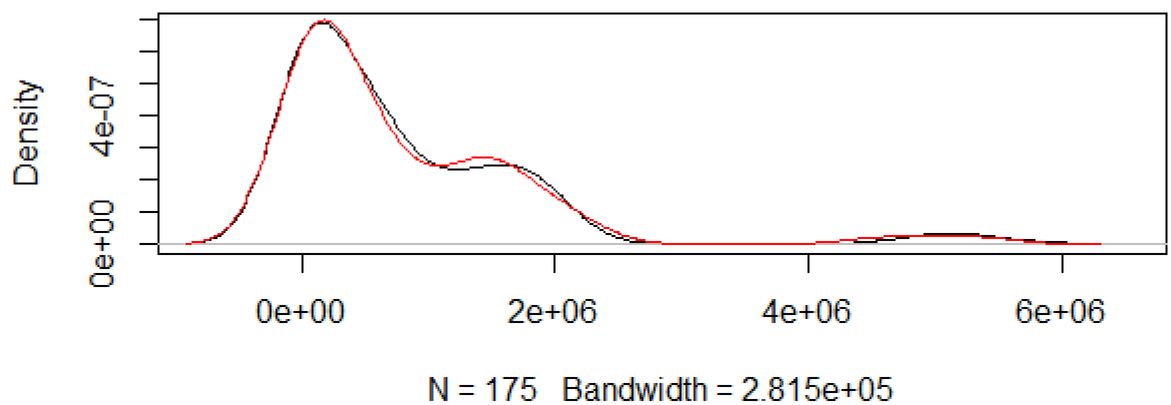
The black curve refers to the observed values of the dependent variable and the red curve refers to the predicted values of the same based on column (3) of Table 5.

Figure 8: Kernel Density comparison for measles



The black curve refers to the observed values of the dependent variable and the red curve refers to the predicted values of the same based on column (3) of Table 6.

Figure 9: Kernel Density comparison for TT



The black curve refers to the observed values of the dependent variable and the red curve refers to the predicted values of the same based on column (3) of Table 7.

Table: 8 Sample-splitting Results (Regression coefficients for test set after 'selecting' predictors from training set) using LARS

	BCG	DPT	OPV	Measles	TT
Funds	-.0000314 (0.944)	.0006235 (0.155)	.0004215 (0.459)	—	.0003433 (0.515)
Infrastructure	74.57* (0.069)	120.88*** (0.005)	171.76*** (0.001)	166.18** (0.012)	242.8649*** (0.000)
SDP	.7006* (0.052)	.5925 (0.129)	.83139* (0.082)	698195 (0.141)	1.623716* (0.027)
Population	.1151*** (0.000)	.10048*** (0.000)	.084735*** (0.000)	.0840251*** (0.000)	.0838086*** (0.000)
Trend	-42788.9** (0.010)	-54091.07*** (0.001)	-72880.76*** (0.001)	-26855.44 (0.156)	-42018.85** (0.028)
Funds*infr	1.19e-06*** (0.000)	9.54e-07** (0.013)	1.26e-06*** (0.001)	1.12e-06*** (0.001)	1.00e-06** (0.049)
$Fund^2$	—	—	—	—	—
NE	—	—	—	—	—
Bimaru	—	6654.919 (0.967)	51022.5 (0.765)	—	-66388.92 (0.791)
NE*funds	—	-.0012305*** (0.005)	-.001348** (0.010)	-.0009687*** (0.000)	—
Bimaru*funds	-.0008565* (0.066)	-.0005129 (0.644)	-.0009324 (0.421)	—	.000701 (0.672)
constant	189580.5 (0.026)	246896.9 (0.006)	334020.5 (0.003)	128898 (0.184)	119873.7 (0.249)
R^2	0.9906	0.9884	0.9841	0.9772	0.9832

The numbers in parenthesis refers to p-values.
*, **, *** indicates 10%, 5%, 1% significance level respectively.
Heteroscedasticity corrected robust standard errors have been used

Table: 9 Sample-splitting Results (Regression coefficients for test set after 'selecting' predictors from training set) using LASSO

Variable	BCG	DPT	OPV	Measles	TT
infrastructure	94.23** (0.0044)	133.2** (0.000209)	184.2** (2.25e-05)	161.1*** (0.0003)	81.76* (0.0826)
Funds*infrastructure	9.872e-07*** (<2e-16)	9.089e-07*** (4.1e-14)	1.075e-06*** (6.09e-14)	1.104e-06*** (1.56e-13)	1.477e-06*** (<2e-16)
Population	0.111*** (<2e-16)	0.1016*** (< 2e-16)	0.08672** (< 2e-16)	0.08458*** (< 2e-16)	2.931e-02*** (<2e-16)
Constant	1.280e+04 (0.4851)	7.498e+03 (0.701596)	7.293e+03 (0.755)	2.365e+04 (0.336462)	-6.240e+03 (0.7748)
R^2	0.9881	0.9852	0.9784	0.9748	0.9829
n	90	90	90	90	90

The numbers in parenthesis refers to p-values.
*, **, *** indicates 10%, 5%, 1% significance level respectively.
Heteroscedasticity corrected robust standard errors have been used

Table 10: Sample-splitting results (Sign and significance of regression coefficients for test set after 'selecting' predictors from training set) using LASSO from the full set of covariates.

Variable	BCG	DPT	OPV	Measles	TT
Infrastructure	+ ***	+*	+***	+	+***
Population	+ ***	+***	+***	+***	+***
Bihar	X	-	X	-***	X
Bihar*funds	-*	-	-***	-***	-***
Bihar*infra	X	X	-***	+***	+***
Bihar*funds*infra	X	+	X	X	+***
MP	+	+	+**	+	+
MP*funds	-	X	X	X	X
MP*infra	X	-	-**	-	-
MP*fund*infra	+	X	X	X	X
Rajasthan	-	-	+	-	+***
UP	+ ***	+***	+***	-***	-
UP*funds	X	X	X	+***	+***
UP*infra	-***	-***	-***	X	X
UP*funds*infra	X	X	X	X	-***
Jharkhand	+	X	X	X	X
Jharkhand*infra	X	X	X	X	+
AP	+	+*	X	+*	+
AP*funds	-	-	+*	-	X
AP*infra	X	X	X	X	-
AP*fund*infra	+	X	-*	X	X
WB	+	X	X	X	X
WB*funds	+	X	X	X	X
WB*fund*infra	-	X	X	X	X
Karnataka	X	X	X	+	X
Karnataka*infra	-	X	X	-	X
Chhattisgarh	X	X	X	X	+
Chhattisgarh*infra	X	X	X	X	-
Gujarat	X	X	X	X	-
Gujarat*funds	X	X	X	X	-
Gujarat*infra	X	X	X	X	+
Haryana	X	X	X	X	+
Haryana*infra	X	X	X	X	-
Kerala	X	X	X	X	+
Kerala*infra	X	X	X	X	-
R^2	0.99	0.99	0.99	0.99	0.99
n	90	90	90	90	90