Bayesian synthetic prediction of state level poverty using Indian Household Consumer Expenditure Survey Data

Soumojit Das[1], Atanushasan Basu[2], Partha Lahiri[3], and Shreya Sengupta[4]

[1]University of Maryland, College Park, MD, USA, *soumojit@umd.edu*

[2]Deputy Director General, Department of Commerce, Government of India, *atanushasan@yahoo.co.in*

[3]University of Maryland, College Park, MD, USA, *plahiri@umd.edu*

[4]National Sample Survey Organization, Government of India, *sengupta.shreya@yahoo.com*

Abstract

Goal 1 of the 2030 Agenda for Sustainable Development, adopted by all United Nations member States in 2015, is to end poverty in all forms everywhere. The major indicator to monitor the goal is the so-called headcount ratio or poverty rate, i.e., proportion or percentage of people under poverty. In India, where nearly a quarter of population still live below the poverty line, monitoring of poverty needs greater attention, more frequently at shorter intervals (e.g., every year) to evaluate the effectiveness of planning, programs and actions taken by the governments to eradicate poverty. Poverty rate computation for India depends on two basic ingredients – rural and urban poverty lines for different states and union territories and average Monthly Per-capita Consumer Expenditure (MPCE). While MPCE can be obtained every year, usually from the Consumer Expenditure Survey on shorter schedules with a few exceptions where the information is obtained from another survey, determination of poverty lines is a highly complex, costly and time-consuming process. Poverty lines are essentially determined by a panel of experts who draws their conclusions partly based on their subjective

opinions and partly based on data from multiple sources. The main data source the panel uses is the Consumer Expenditure Survey data with a detailed schedule, which are usually available every five years or so.

In this paper, we undertake a feasibility study to explore if estimates of headcount ratios or Poverty Ratios in intervening years can be provided in absence of poverty lines by relating poverty ratios with average MPCE through a statistical model. Then we can use the fitted model to predict poverty rates for intervening years based on average MPCE. We explore a few in this work models using Bayesian methodology. The reason behind calling this 'synthetic prediction' rests on the synthetic assumption of model invariance over years, often used in the small area literature. While the data-based assessment of our Bayesian synthetic prediction procedure is encouraging, there is a great potential for improvements on the models presented in this paper, e.g., by incorporating more auxiliary data as they become available. In any case, we expect our preliminary work in this important area will encourage researchers to think about statistical modeling as a possible way to at least partially solve a problem for which no objective solution is currently available.

Keywords: Poverty; Sustainable Development Goals; Consumer Expenditure Surveys; Data Integration; Synthetic prediction; Bayesian Beta Regression

1 Introduction

Poverty and unemployment are the prime concerning factors affecting socio-economic progress of human society all over the world throughout history, especially in developing and lesser developing countries. Every country or government needs reliable information on these aspects for effective reparation to ensure peaceful living and equal opportunity for their citizens. Though standard concepts, definitions, and measures are available pertaining to unemployment, the Indian government has not yet decided on a methodological definition of poverty measurement. It is very difficult to define and determine objectively what constitutes poverty for its multidimensional, subjective, judgmental and relative nature.

[0]The views expressed in this paper are exclusively of the authors and not of the organizations they belong.

Standardization of concepts and definitions are a primordial necessity to obtain estimates on important socio-economic indicators within a definite geographical boundary. But the concept of 'poverty' is not unique all over the globe and is even changing over time. This challenge motivated the authors to undertake such a study. The concept of poverty varies within and across countries, over time, among various socio-economic groups and even from individual to individual. Nevertheless, each country tries to assess its gravity to formulate policies for providing with all of its citizens certain reasonable or minimal acceptable standard of living. The methods of estimating the poverty in numerical terms by different countries may differ, although mostly it is based on consumption or income of individuals or of a family in monetary terms.

At the 3^{rd} Pakistan statistical conference at Lahore in February 1956, Professor P.C. Mahalanobis noted: "The solution of these twin problems of poverty and unemployment will require much hard thinking and positive action based on factual information." It is the responsibility of statisticians to provide reliable estimates of these parameters to the policymakers. This will help the nation and the human society at large to effectively find suitable remedies.

United Nations Development Program (UNDP) observed, "Traditionally, and still in many circles; well-being is understood as material progress measured by income. According to this view, countries worthy of receiving external financial assistance are those below a certain arbitrary level of income. And poverty is measured by counting the number of people living under an arbitrary poverty line." A global indicator framework was developed by the Inter-Agency and Expert Group on Sustainable Development Goal (SDG) Indicators (IAEG-SDGs) and agreed upon, including refinements on several indicators, at the 48^{th} session of the United Nations Statistical Commission held in March 2017. The global indicator framework was later adopted by the General Assembly on 6 July 2017 and is contained in the Resolution adopted by the General Assembly on Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development. Assigning utmost importance to eradication of poverty, Goal 1 of the SDG is to end poverty in all its forms

everywhere by 2030.

In India, around a quarter of the population still live below the poverty line [2], [3]. Thus monitoring of poverty needs greater attention, more frequently at shorter interval, like every year, to evaluate the effectiveness of planning, programs and actions taken by the governments to eradicate poverty. But the lack of available detailed data in each year is the main impediment to undertake such exercise. India with a population of more than 1.25 billion people needs massive resources in terms of cost and time to conduct a detailed household survey covering the entire country for collecting information on consumption expenditure and then to develop a suitable methodology to estimate poverty. Note that, this is in reference to the resource requirement to conduct pan-India sample survey on household consumer expenditure based on detailed schedule of inquiry. The approximate population is a Census projection estimate based on the NSSO data [10], [9]. Expense of such huge amount of resources is almost a luxury for any developing or a lesser developed nation, considering other prioritized vital socio-economic needs of their people.

The complexities in defining and estimating poverty reliably (pertaining to how the Indian Government does it) along with related involvement of huge resources motivated us to undertake the present study by finding an appropriate statistical relationship between related socio-economic variables like household consumption expenditure, unemployment, and poverty. If poverty, in terms of proportion or percentage living below a poverty line during a period of reference, can be estimated by such a relationship, then there could be a huge savings of resources as explained above. In India, the detailed survey on household consumer expenditure is generally done once in every five years. This study can be further extended to estimate poverty in the intervening years on the basis of information on household consumer expenditure collected in the intervening years on a much shorter schedule.

The main objective of the paper is to encourage researchers to consider modeling as an alternative to the highly expensive and time-consuming approach that is currently implemented roughly every five years. With the suggested modeling approach, poverty rate

estimates can be produced more frequently than is currently done. We did not mean to advocate any specific model, as models can always be improved. As a limited evaluation, we have taken a Bayesian regression approach, which provides a reasonable solution and offers flexibility in various inferences compared to the regression implemented using a classical method especially when the sample size is relatively small, as is in our case (17 data points).

The rest of this paper is organized in the following way. Section 2 discusses how poverty is estimated in India. Section 3 provides details of the empirical evaluation of a few statistical models. Finally, we present the concluding remarks and discuss how this work can be extended in the future.

2 Poverty estimation in India

In India, poverty is estimated in two steps. Firstly, for each state and union territory (UT), the poverty line is estimated separately for rural and urban areas in terms of monthly/daily per-capita expenditure required to maintain a minimal standard of living. The expenditure requirement to maintain a minimum standard of living is determined distinctly for different selected categories of food and non-food commodities using consumption expenditure data on food and non-food items, price relatives from the National Sample Survey (NSSO), Central Statistical Office (CSO). Minimal nutritional requirement norm from medical expert bodies like Indian Council of Medical Research (ICMR), etc. and other related data compiled by different government organizations, according to methodologies adopted by Expert Groups, were constituted by the Government from time to time to estimate poverty. The non-food item basket mainly comprises clothing, footwear, education, rent, conveyance and medical expenses.

After deciding on the poverty line, the proportion or Percentage of Population (PoP) with average monthly per-capita consumer expenditure (MPCE) below this poverty line is calculated. Over the years, researchers have been using the term PoP interchangeably with

Head Count Ratio (HCR) or Poverty Ratio (PR). In order to avoid any confusion, we use PoP throughout the paper. The distribution of population by MPCE is available from the results of the National Sample Survey Office (NSSO) surveys.

The MPCE is obtained from the results of detailed survey on Household Consumer Expenditure conducted every five years in general, by NSSO of the Government of India. The National Sample Survey Office (NSSO) usually conducts a detailed survey on Household Consumer Expenditure every five years. However, if necessitated by various stakeholders, the detailed survey may also be conducted in the intervening years subject to the decision by the competent authority of Government in this regard. MPCE is also available from the results of various other socio-economic surveys conducted by NSSO in each year, based on five shot or one shot query in the schedule. Sometimes, information on Household Consumer Expenditure is also collected in the intermediate years with a much shorter schedule, if the Government decides so. In summary, MPCE estimates are available every year either from the Household Consumer Expenditure survey (with short or detailed schedule) or some other surveys. For our paper, we used MPCE derived from Household Consumer Expenditure surveys with detailed schedule that were conducted in 2009-2010 and 2011-12. Information of expenditure on around three hundred and fifty food and non-food items are generally collected through these detailed surveys. The recall periods for different item groups are different and varies from Round to Round. A Round is a specific time period, normally one year or six months, during which surveys on assigned subjects are conducted by NSSO. For example, the last detailed survey on Household Consumer Expenditure was conducted in 75^{th} Round during 2017-2018 by NSSO.

The monthly consumer expenditure is computed at first at household level with reference to a combination of recall periods for various item groups. After that, MPCE, at the individual level, is derived by dividing the monthly consumer expenditure by the size or number of household members in the selected household. Thus, MPCE is assumed to be the same for the all the household members in a household. Generally, the MPCE is calculated by three approaches, namely Uniform Reference Period (URP), Mixed Reference Period

(MRP) and Modified Mixed Reference Period (MMRP). In URP, household consumer expenditure is recorded for a uniform recall period of last 30 days preceding the date of survey on each item. In MRP, the recall periods are the last 365 days on items like clothing and bedding, footwear, education, institutional medical care, and durable goods and the last 30 days on remaining items. In MMRP, the recall periods are last 7 days on items like edible oil, egg, fish and meat, vegetables, fruits, spices, beverages, refreshments, processed food, pan (betel leaves), tobacco and other intoxicants; last 365 days on items like clothing and bedding, footwear, education, institutional medical care, and durable goods and last 30 days on remaining items.

Since the primary objective of this paper is to develop a model to predict the PoP on private consumption expenditure data, detailed discussion on successive methodologies adopted to estimate poverty line in India is beyond the scope of this paper. However, the way methodologies evolved in India over time to estimate poverty line and poverty ratio are presented in brief, chronologically, in the following paragraphs.

Historically, attempts to estimate poverty in India can be dated back to as early as 1962. At the time, the Planning Commission constituted a Working Group to develop a suitable methodology for estimating poverty. The expert group submitted its report in the same year. After that, in 1977, a Task force was appointed by the Government under the chairmanship of Dr. Y. K. Alagh. The report submitted by this task force in 1979 was accepted by the Government. The methodology was further improved by the Expert Group chaired by Professor D. T. Lakdawala constituted in 1989. The Group submitted its report in 1993, which was accepted by the then Planning Commission in 1997. An Expert Group under Suresh D. Tendulkar was constituted in the year 2005, which submitted its report in 2009. The report was accepted by the Planning Commission in 2011. Further improvisation in the methodology was made by another Expert Group constituted in 2012 under the chairmanship of Dr. C. Rangarajan. The Group submitted the report in 2014.

For more details we refer to the following reports: [3], [2], [10], [9].

3 Empirical Evaluation

In this section, we empirically evaluate a few models for prediction of the percentage of population under the poverty line (PoP) using just one auxiliary variable, average monthly per-capita consumer expenditure or MPCE. To this end, we consider household consumer expenditure (HCE) surveys for the years 2009-10 and 2011-12, conducted by the National Sample Survey Organization (NSSO), during its 66th Round (2009-10) and 68th Round (2011-12), respectively. There are several reasons for selecting these two surveys for our empirical work. First, these two are the latest detailed surveys on household consumer expenditure for which results are available in the public domain. Secondly, PoP estimates, based on these surveys, are also available for these periods. Let us elaborate on this a bit. The Household Consumer Expenditure (HCE) survey with detailed schedule for the years 2009-10 and 2011-12 are the only recent surveys that produced both PoP and MPCE. This allows us to fit a model using 2009-10 survey data and then evaluate the estimates based on the fitted model and MPCE for 2011-12 because PoP values are also available for 2011-12. Another HCE survey with detailed schedule was conducted in 2017-18, but unfortunately estimates from that survey are yet to be made publicly available. The earlier quinquennial result, based on a detailed survey, was available for the year 2004-2005. However, the methodology for measurement of poverty in India was revised by the Expert Group under the Chairmanship of Dr. C. Rangarajan in the year 2014 using 2009-10 and 2011-12 Household Consumer Expenditure data. These are the reason that motivated us to choose 2009-10 and 2011-12 data. We have used the poverty estimates based on methodologies recommended by Tendulkar Committee with mixed reference period (MRP). Moreover, the coverage of items, sampling designs, schedules of inquiry, and concepts and definitions are similar for these years.

For this study, we consider $n = 17$ major Indian states with a combined share of more than 90% of total Indian population. Moreover, for these states the relative standard errors of estimates of average MPCE are well below 5% for the year 2011-12, so we could avoid

complex models that incorporate sampling variability.

We are calling our method Bayesian synthetic prediction. Let us take this opportunity to elaborate on this a little. Basically, the idea of the paper is to produce the PoP values for a certain year using a model fitted using data on both PoP and MPCE values available for a most recent (past) year and using MPCE of the present year as a predictor variable, which is available more often (every year). Following small area literature [11], we are using the term synthetic because this rests on a synthetic assumption of invariance of the model (estimated using a recent year) over years. We think our problem is more like a prediction, rather than estimation, because the PoP values would be unavailable for the year of interest. We chose the year 2011-12 for evaluation because PoP values are available to test out the proposed method.

Let p_{ij} denote the proportion of population under the poverty line (PoP) and x_{ij} denote the average monthly per capita consumer expenditure (MPCE) for the ith state in the jth year, $i = 1, \cdots, 17;\ j = 1, \cdots, J$. For our data analysis, $j = 1$ refers to the base year 2009-10 when data $(p_{i1}, x_{i1},\ i = 1, \cdots, 17)$ is available to fit a model establishing relationship between p_{i1} and $x_{i1},\ i = 1, \cdots, 17$; $j = 2$ refers to the evaluation year 2011-12 when we use x_{i2} to predict p_{i2} using the fitted model based on the base year and evaluate based on the observed $p_{i2},\ i = 1, \cdots, 17$.

Since p_i are proportions lying on a continuous scale $(0, 1)$, Beta regression could be a reasonable way to model p_i. Beta regression [4] uses the beta distribution as the likelihood of the data (in contrast to the normal distribution in the case of usual regression). To elaborate, we assume p_i's are independent with the following Beta density:

$$f(p_i|a_i, b_i) = \frac{p_i^{a_i-1}(1-p_i)^{b_i-1}}{B(a_i, b_i)},$$

where $B(.,.)$ is the usual Beta function. The parameters for the Beta distribution, i.e., a_i

and b_i, are given by

$$
\begin{aligned}
a_i &= \mu_i \cdot \phi, \\
b_i &= (1 - \mu_i) \cdot \phi, \\
g(\mu_i) &= x_i'\beta, \quad i = 1, \cdots, n,
\end{aligned}
$$

where $g(.)$ is a known link function linking the mean of the Beta distribution to a $p \times 1$ vector of auxiliary variables x_i; β is a $p \times 1$ vector of unknown regression coefficients and ϕ is an unknown scale parameter. It is possible to generalize the model by assuming the scale parameter ϕ to be state specific, i.e. ϕ_i, explainable by a vector of state specific auxiliary variables. But we do not pursue this generalized model in this paper.

We implement a fully Bayesian analysis, which requires specifying prior distributions for all model parameters. In our application, we have only one auxiliary variable – (average) MPCE. Thus, we have three unknown parameters of the Beta regression model – intercept, slope and the scale parameter. We assume weakly informative priors for the model parameters – a $N(0, 10^2)$ prior on the intercept, a $N(0, 2.5^2)$ prior on the slope and an exponential distribution with mean 1 prior on the scale parameter ϕ.

To fit the Beta regression model, we use the **Stan** [12] probabilistic programming language with the statistical software **R**. **Stan** has many inbuilt options that can handle different types of priors on different parameters of the model. Using `stan_betareg`, these prior distributions of preference can be set using the `prior_intercept, prior` and `prior_phi` arguments. For detailed information, we refer to the `rstanarm` paper, [7]. Since our objective for this work was not to do a comparative study on prior selection for the Bayesian models, rather we want to motivate modeling as an approach to study poverty, thus we keep all the prior choices default to the **Stan** and `rstanarm`.

The choice of the link function $g()$ is important. We consider two widely used link functions: 'logit' and 'probit'. Note that, `rstanarm` uses 'logit' link by default for beta regressions. Still, we wanted to try out which link works better for our data. **Stan** and

`rstanarm` conveniently provides Leave-One-Out Cross-Validation (LOO-CV) information criterion to evaluate and select the model that fits the data better. We use the `loo` package [13] with **Stan** to select the link function which works better for our data. In the tables below, we provide the LOO diagnostics for both the regions.

Table 1: Comparison of the two different link functions for the Rural region. We see that Logit link fits our data better as per the LOO criterion.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
logit link	0	0	12.988	0.539
probit link	-0.132	0.050	12.856	0.538

Table 2: Comparison of the two different link functions for the Urban region. We see that Logit link fits our data better as per the LOO criterion.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
logit link	0	0	14.875	0.549
probit link	-0.084	0.010	14.790	0.547

We observe from the tables 1 and 2 that for both the regions, our models fit the data better with the known specification of link function $g()$ as 'logit'. As we have already noted, all the rest of the various prior-tuning options were left default while we were fitting the two different models – we only varied the link functions.

Another way of fitting the data, that we have explored in this work, can be as follows:

1. We can take logit transformation on the p_i's.

2. We can then use these transformed p_i's and fit a linear regression model with the MPCE as the covariate, often referred to as the logistic regression for the continuous covariates.

3. We can do the above in a Bayesian way by using priors on the model parameters. Again, we will not go into the different choices of prior. Rather, we will keep all the prior options as default Weakly Informative Priors (WIP).

Next, we fit the logit transformed p_i's to the MPCE in linear regression setup (logistic regression). We do this in a Bayesian setup by specifying priors. The motivation of doing this in a Bayesian way is to facilitate meaningful model comparison and selection using LOO-Information Criterion. We, again, compare this model with the Bayesian Beta Regression with logit link (since we already established that the Bayesian Beta regression with the logit link performs better than with the probit link). We do this for both the regions.

Table 3: LOO-IC comparison of Bayesian Beta Regression with logit link and Bayesian Linear Regression on logit transformed data for the Rural region.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
Bayesian Beta Regression	0	0	12.988	0.539
Bayesian Linear Regression	-16.409	2.276	-3.421	2.577

Table 4: LOO-IC comparison of Bayesian Beta Regression with logit link and Bayesian Linear Regression on logit transformed data for the urban region.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
Bayesian Beta Regression	0	0	14.875	0.549
Bayesian Linear Regression	-18.835	3.194	-3.960	3.223

From the two tables 3 and 4, it is clear that for both the regions, Bayesian Beta regression with logit link provides the better model fit compared to the Bayesian logistic regression according to the LOO-CV Information criterion.

To further assess the performance of the Beta regression model, we compare the posterior predictive distributions of rural (Figure 1) and urban (Figure 3) areas for the 17 states with that of the observed distribution using the 2009-10 data. The idea behind posterior predictive checking is simple: if a model is reasonable, then we should be able to use it to generate data that looks a lot like the data we observed. This provides a nice visual check if the model fits the data well.

For further details on the posterior predictive distributions, Bayesian Data Visualization and implementation in **Stan**, we refer to the book by Gelman et al. [6] and the papers [1],

[5], [8].

The posterior predictive checks suggest that the data simulated from the posterior predictive distributions agree well with the actual observed data. For the next model validation, we plot the Bayesian credible intervals in the same graph with the observed PoP for the year 2011-12. We do this for both the regions. The figures are displayed in Figure 2 for the Rural region and Figure 4 for the Urban region. We see for both the regions, all the observed PoP values for the evaluation year 2011-12 lie within the 90% Bayesian credible interval. We also notice that these credible intervals are pretty wide. Availability of more appropriate state-specific covariates and inclusion of such into the model will very likely improve these intervals and the model fits in general.

Table 5: Means (estimates), standard deviations (standard errors), and 6 percentiles of the posterior distribution of parameters (i.e., intercept and slope) of the Beta regression model computed using 2009-10 data for Rural areas; the column denoted by 50% provides medians of the posterior distribution of the parameters ; the lower and upper limits of 95% credible intervals for the parameters are obtained from the columns denoted by 2.5% and 97.5%, respectively. Similarly, we can find the lower and upper limits of 70% credible intervals from the columns denoted by 15% and 85%.

Rural	mean	sd	2.5%	10%	15%	85%	90%	97.5%
(Intercept)	1.1620	0.7470	-0.2754	0.2248	0.4071	1.9298	2.1145	2.6655
MPCE	-0.0017	0.0007	-0.0031	-0.0026	-0.0024	-0.0010	-0.0009	-0.0004

Table 6: Means (estimates), standard deviations (standard errors), and 6 percentiles of the posterior distribution of parameters (i.e., intercept and slope) of the Beta regression model computed using 2009-10 data for Urban areas; the column denoted by 50% provides medians of the posterior distribution of the parameters ; the lower and upper limits of 95% credible intervals for the parameters are obtained from the columns denoted by 2.5% and 97.5%, respectively. Similarly, we can find the lower and upper limits of 70% credible intervals from the columns denoted by 15% and 85%.

Urban	mean	sd	2.5%	10%	15%	85%	90%	97.5%
(Intercept)	0.3198	1.0897	-1.8262	-1.0663	-0.7761	1.4206	1.6854	2.4838
MPCE	-0.0008	0.0006	-0.0019	-0.0015	-0.0013	-0.0002	-0.00002	0.0004

Tables 5 and 6 provide details of Bayesian Beta regression (with logit link) fit by **RStan**.

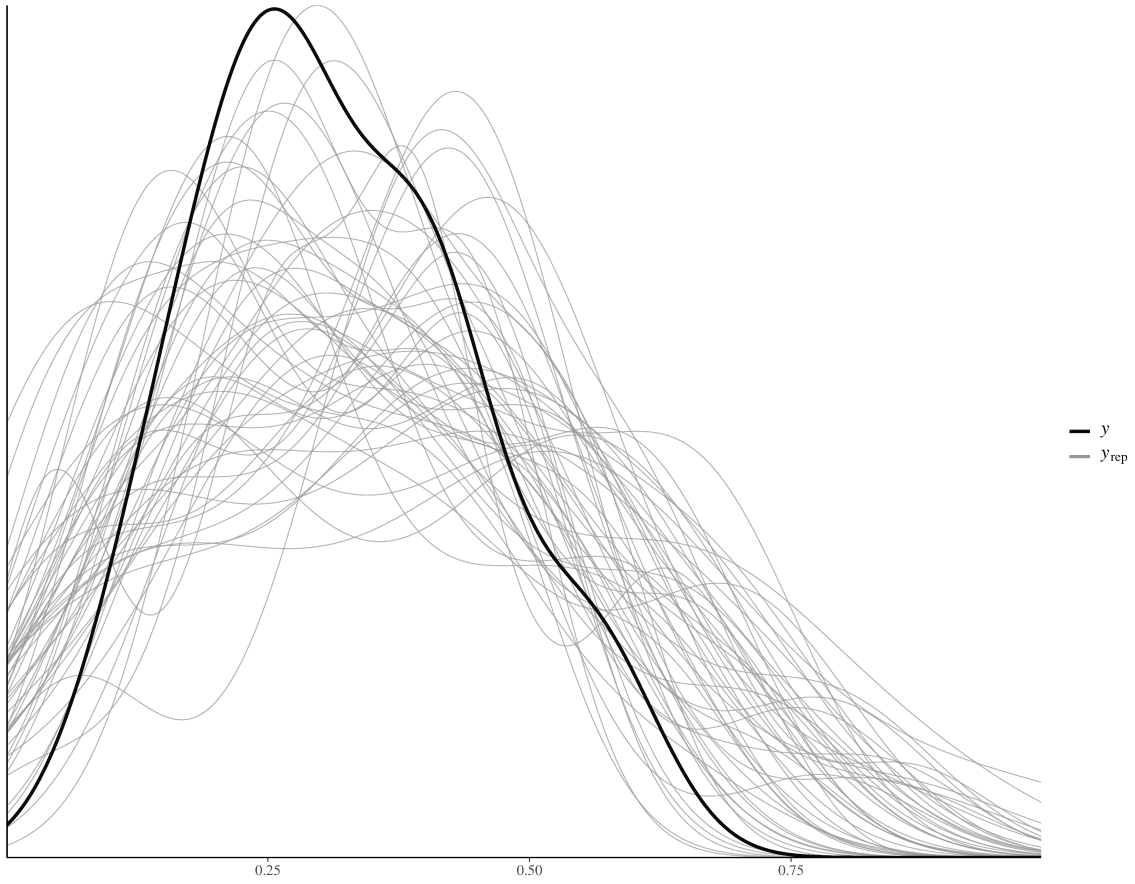

Figure 1: Comparison of the observed distribution of the PoP computed using the 2009-10 data for rural areas in the 17 states (darker curve) and the corresponding distributions constructed using simulated samples from the posterior predictive distribution of the PoP derived from the selected Beta regression model (lighter curves).

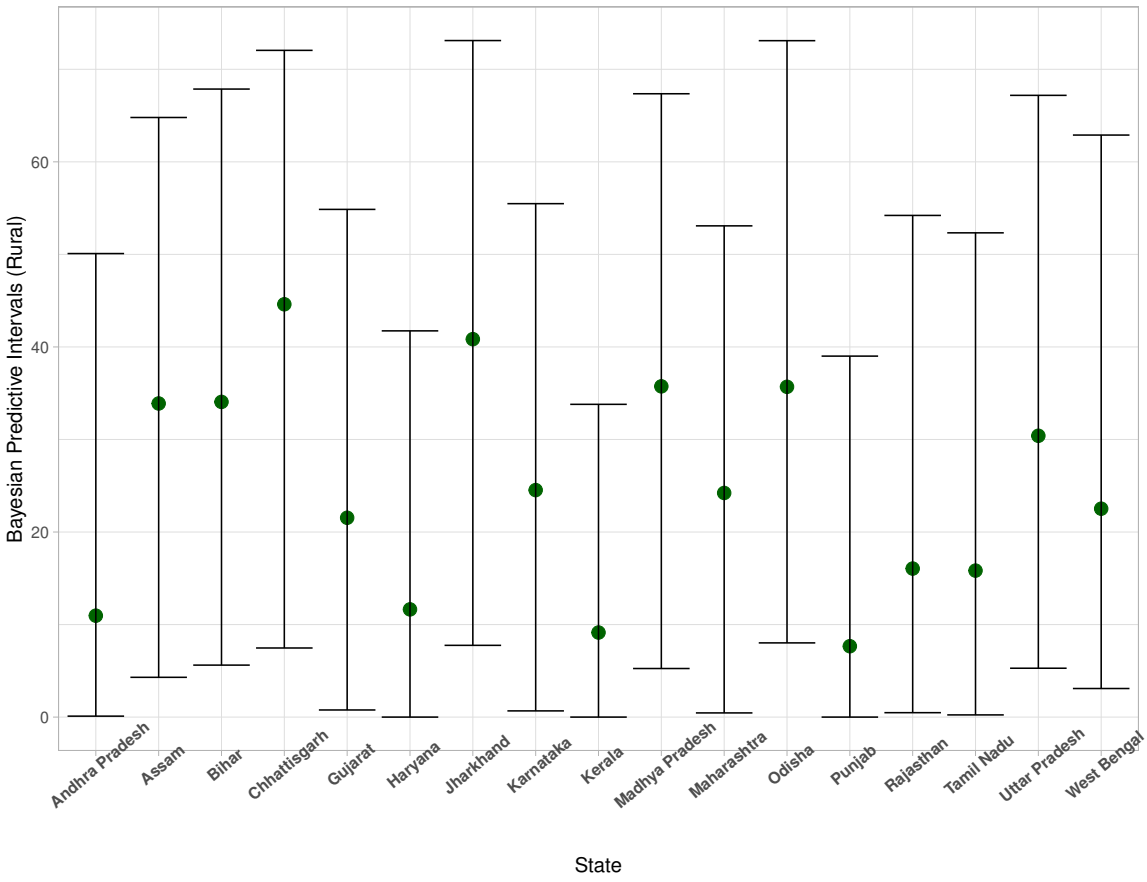

Figure 2: Bayesian predictive intervals (90%) of PoP for rural areas in 17 different states; green dots represent the observed PoP; we can see that all the green dots are contained within the predictive intervals.

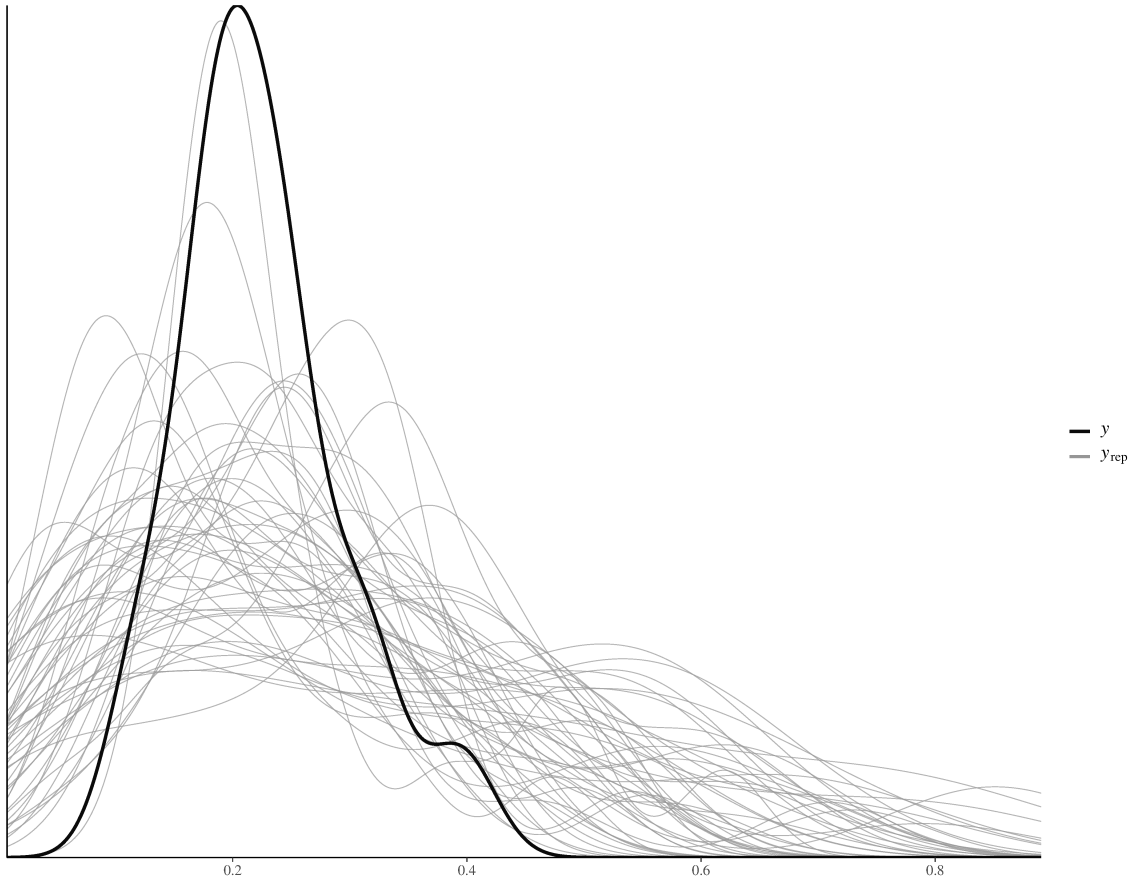

Figure 3: Comparison of the observed distribution of the PoP computed using the 2009-10 data for urban areas in the 17 states (darker curve) and the corresponding distributions constructed using simulated samples from the posterior predictive distribution of the PoP derived from the selected Beta regression model (lighter curves).

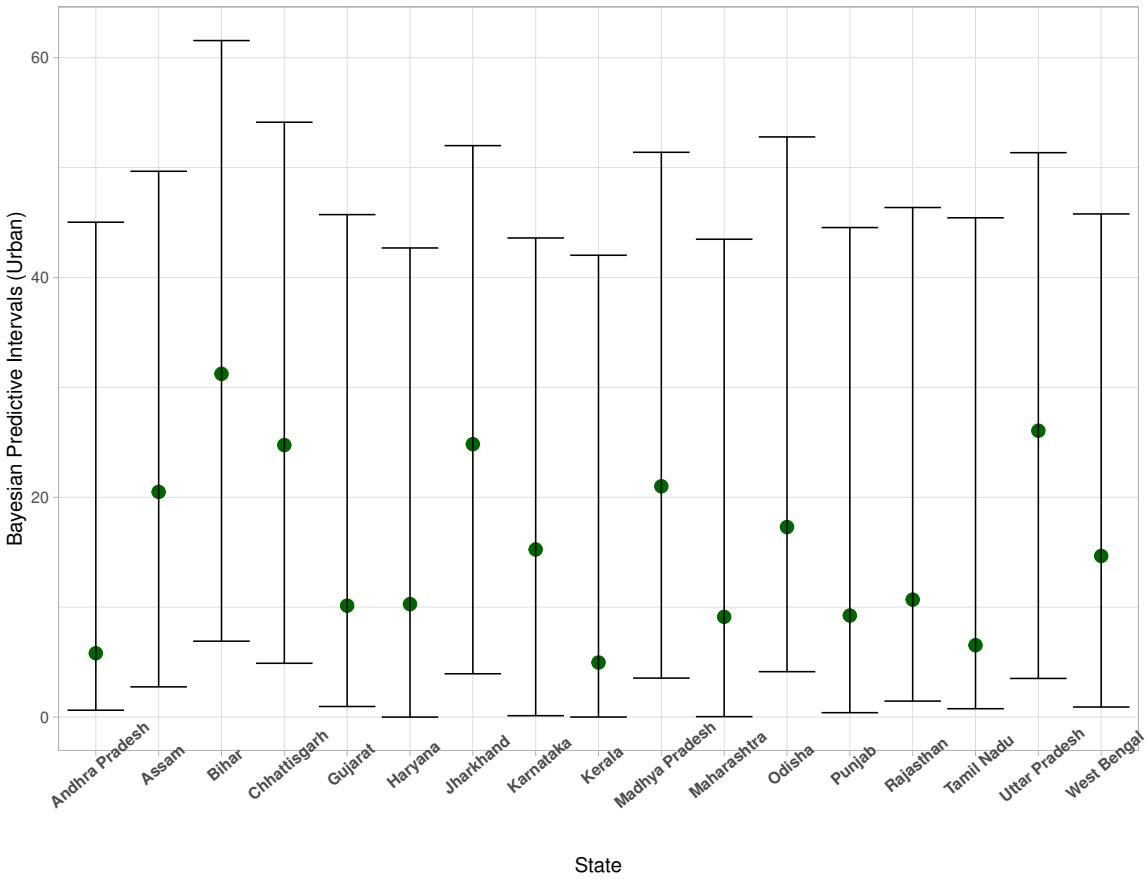

Figure 4: Bayesian predictive intervals (90%) of PoP for urban areas in 17 different states; green dots represent the observed PoP; we can see that all the green dots are contained within the predictive intervals.

From Tables 5 and 6 we see that slope estimates for both rural and urban areas are negative, which makes sense as we can expect PoP to be negatively related to MPCE. We observe for the Rural regions, the inercept and the slope are both significant at 70% because the credible intervals do not include 0. For the Urban regions, the 70% credible interval for intercept contains 0 whereas that for the slope does not, making it significant at that level. We suspect for the Urban regions, where poverty maybe indicated by various other complex factors other than just average MPCE, modeling poverty measures such as the PoP is much more affected by the presence of outliers, inadequate sample size and of course unavailabilty of other possible state-specific auxiliary variables. The availability of such strong state-specific covariates will not only improve the model fit, they may also explain the outliers better. For the Rural regions however, we see an overall better fit that that of the Urban regions. This may be indicated by simpler lifestyle in the rural regions and as a result average MPCE explains quite a lot. Nevertheless, in presence of more relevant state-specific covariates, more light may be shed on better understanding of how poverty depends on other factors that just expenditure measurements such as average MPCE. Future work in this direction should be. For evaluation, we use the models for the 2 regions to predict the PoP values for the year 2011-12. We use the estimates from our Bayesian Beta regression models, to get the predicted values for the year 2011-12 for both Rural and Urban regions. Since we already have the original PoP values for year 2011-12, we used them to measure how good our models could predict the observed values.

Table 7 compares the prediction errors of the Bayesian Beta regression for rural and urban areas. We define the prediction errors simply as the difference between predicted values and the observed (true) values; which are actually available for the year 2011-12. Prediction Error $= \hat{\text{PoP}}_{predicted} - \text{PoP}_{observed}$. We prefer this prediction error over absolute or squared prediction error because this will give us an idea about underestimation or overestimation.

Table 7: Summary statistics of *prediction errors* (in predicting the values for the year 2011-12) for the two regions.

	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
Rural	-9.0174	-5.4190	-2.7386	-2.7804	0.3135	3.5128
Urban	-3.4669	-0.6533	1.5886	2.4786	6.2117	9.4941

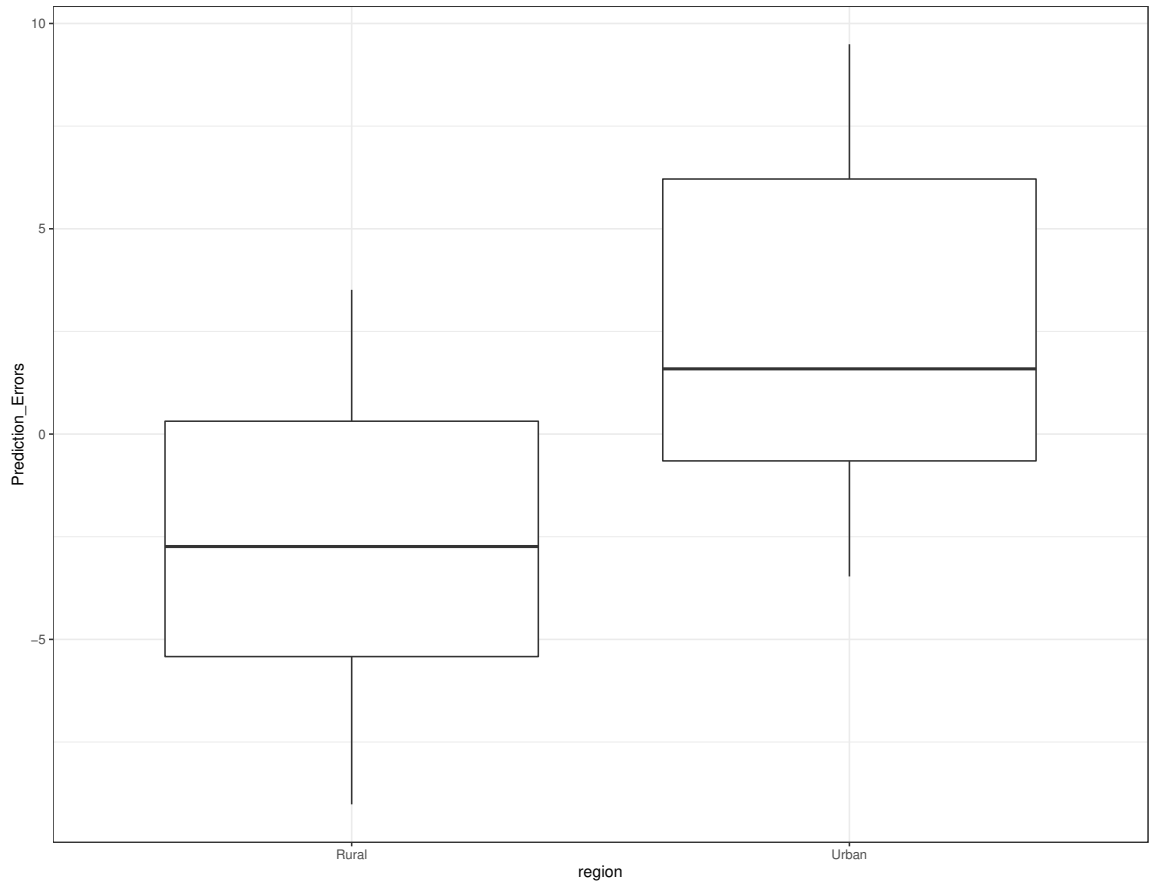

Figure 5: Boxplots of *Prediction Errors* (in predicting the values for the year 2011-12) for the 2 regions. Prediction errors are simply defined as the difference between the predicted values (for 2011-12) and the observed values (for 2011-12).

From Table 7 and Figure 5 we see that our predicted values underestimate for the Rural regions whereas overestimates for the Urban regions. We think that with better availability of some state-specific covariates will solve this problem and improve the model fits overall.

Finally, we would like to add that, while the Bayesian Beta regression with just one state level covariate MPCE is doing a reasonable job in our limited study, it is possible to improve on the model by incorporating more state level covariates and also spatial correlations among different states that may remain even after incorporating relevant state level covariates.

4 Discussion and Conclusion

The data we used here are estimates obtained from sample surveys conducted by the Indian National Sample Survey organization. We investigated a modeling approach to estimate poverty for different states in India. We explored a few possible models, gave some ideas on model selection, validation and finally used the best model (using a Bayesian LOO-CV information criterion) to predict the PoP values. We observe some outliers in this preliminary investigation, which may be reduced by including more state specific auxiliary variable(s). For example, auxiliary variables such as average monthly per-capita expenditure (MPCE) on different categories of items (e.g., food and non-food), indicators of inequality in MPCE distribution like Gini's coefficient and unemployment rates or labor force participation ratios at state levels can well be explored. Inclusion of a sizable number of relevant auxiliary variables requires data from more states. However, estimates from the small states are subjected to large standard errors. Hierarchical Bayesian Beta regression model can be explored in the future for incorporating sampling errors of the estimates and inclusion of data over time when such data becomes available.

One thing to note here that, this paper was not able to contribute to the recent debate on the absence of poverty estimates in India and the comparison of poverty estimates based on various methods. The reasons being, when we started this study, we had hoped to provide an estimate of PoP or the headcount ratio rather than attempting to estimate the poverty line (based on which such proportions were obtained) itself. This is a limitation of our work here that we fail to define a poverty threshold. On the other hand, what we try to do here instead is to give an estimate of the percentage of people below a set poverty threshold,

which is determined by an expert panel every 5-10 years. In the next few lines, we will try to address why we undertook such a study instead of attempting to estimate the poverty line itself. In India, post independence, the development of methodologies for estimating poverty has evolved over time. Several committees were established by the Government, e.g, Alagh Committee (1977), Lakdawala Committee (1989), Tendulkar Committee (2005) and C. Rangarajan Committee (2012) to recommend the poverty threshold. The headcount ratios are then obtained from average MPCE from the quinquennial detailed NSSO surveys on household consumer expenditure. The ever-changing socio-economic pattern and its dynamics necessitate a new definition of poverty and determination of poverty line afresh almost every time whenever the new data on household consumer expenditure is available based on NSSO quinquennial surveys. The determination of poverty lines each time not only requires NSSO data but data from various other sources as well, both at individual and aggregate level. Moreover, there is always an element of subjectivity in defining poverty relative to place and time, because of changing socio-economic scenario. This paper used aggregate data, mainly from NSSO, which is available in public domain, for model based prediction of proportion of population below poverty line using average MPCE. The average MPCE figures from NSSO surveys with short schedule of inquiry are also available for some intervening years between two consecutive quinquennial surveys (which are based on a longer and more detailed schedule). Considering the huge requirement of additional data from sources other than NSSO and the intricacies of deriving methodologies for defining and estimating poverty, we did not attempt the model based estimation of poverty threshold and then getting direct estimates of PoP based on this threshold for this present work. As such, this is a limitation indeed that we fail to estimate such poverty threshold, this is a different study altogether and can be attempted by interested researchers in the future. Our study focuses on giving a reliable estimate of proportion of impoverished people in absence of the detailed schedule and expert recommendation of the poverty line (which is then used by the Government to obtain such proportions) for the intervening years, using the average MPCE. We hope that such proportion estimates for the rural and urban regions

in the intermediate years, where no such direct estimates are available, would facilitate the Government to formulate policies towards poverty eradication in an effective and optimal manner.

While the model found in this paper can potentially be improved in the future, we hope this preliminary work sheds some light on what can be achieved for a problem for which we do not have an existing solution and will encourage researchers to think of other possible alternative solution to this important problem related to Sustainable Development Goals indicators.

Acknowledgements

The research of the first and third author was supported in part by the U.S. National Science Foundation Grant Number SES-1758808 awarded to Partha Lahiri. Comments and remarks made by the two anonymous referees were extremely helpful in re-vising the paper in many ways. We are grateful for those comments and would like to extend our gratitude to the referees. Finally, we would like to thank the editor of the paper for his comments and remarks in finalizing the paper and the form that it is being presented now.

References

[1] PC Buerkner. Brms: Bayesian regression models using stan (r package version 1.6. 1), 2017.

[2] Planning Commission. Report of the expert group to review the methodology for estimation of poverty. Nov 2009.

[3] Planning Commission et al. Report of the expert group to review the methodology for measurement of poverty. *Government of India, New Delhi*, 2014.

[4] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

[5] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019.

[6] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[7] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2018. R package version 2.17.4.

[8] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via stan. *rstanarm: Bayesian applied regression modeling via Stan*, 2018.

[9] India NSSO. Household consumer expenditure across socio-economic group, nss 68th round. July 2012.

[10] India NSSO. Household consumer expenditure across socio-economic groups, nss 66th round. Oct 2012.

[11] John NK Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015.

[12] Stan Development Team. RStan: the R interface to Stan, 2018. R package version 2.17.3.

[13] Aki Vehtari, Jonah Gabry, Yuling Yao, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models. *R package version*, 2(0):1003, 2018.