District-level Estimates of Childhood Malnutrition in India: Application of Small Area Estimation Technique Using Census and DHS Data

Swati Srivastava, PhD Scholar

International Institute for Population Sciences, Mumbai

Email ID-sswati146@gmail.com

Abstract

The Indian Health Service functions under a decentralized approach; however, the lack of district level statistics implies that local authorities are faced with difficulties in making policy decisions without relevant statistics. The Indian Demographic and Health Surveys provide a range of invaluable data at the regional/ national level; they cannot be used directly to produce reliable district-level estimates due to small sample sizes. This study uses small area estimation techniques to derive model-based district-level estimates of childhood malnutrition in India by linking data from the 1998 IDHS and the 2000 Indian Census. The models indicate considerable variability in the estimates of stunting, wasting and underweight across the districts of India. The diagnostic measures indicate that the model-based estimates are reliable and representative of the district to which they belong.

Extended Abstract

Introduction

In recent scenario estimation of any population characteristics (like TFR, IMR, fever etc.) for lower level (subnational, district, zip code etc.) became the vital issues in demographic surveys. The purpose of using such kind of statistical tools is to determine the small area estimates of health-related indicators to make some decentralized approaches, which will helpful to make planning and to allot the resources to needy individuals by triggering out the exact locations. In developing county like India, considerable small area estimates are not available, except to census data. However, the census data is restricted to provide information on socioeconomic and population related indicators and very limited to health indicators. Moreover, the population censuses in India has conducted in every 10 years.

In contrast, cross-sectional surveys, such as the Demographic and Health Surveys (DHS), have become more regular and they collect a substantial amount of nationally representative data. Till date, India had run four rounds of DHS in the years 1992-93, 1998-99, 2005-06 and 2015-16.

Except DHS-4 (2015-16), all the DHSs provided the national as well as state level estimates of demographic and health indicators but cannot be representative of district level estimates because of owing small sample size and high sample variability (Rao, 2014)

Data and methodology

Data source and variables

The purpose of the study is to derive the district level (here area is districts) estimates of childhood malnutrition (stunting, wasting and underweight) across all districts of India for the year 1998-99. To determine area level estimates two types of variables are required. 1) outcome variable, for which small area estimates are required and 2) covariates, which is known for the entire populations and works as the auxiliary information for small area estimates.

In present study outcome variables are childhood stunting, wasting and underweight, which were taken from survey data. The NFHSs is the repeated cross-sectional surveys which were aimed to collect the detail information about the several demographic and health indicators in India. The first NFHS was conducted in 1992-93 to create a pave for important demographic and health related database in India. However, the second NFHS was undertaken in 1998-99 to strengthen the data base and also facilitate implementation and monitoring of population and health programmes in the country. The principal objective of NFHS-2 was to provide state and national estimates of fertility, the practice of family planning, infant and child mortality, maternal and child health, and the utilization of health services provided to mothers and children. NFHS-2 adopts the multi-stages stratified sampling technique in rural and urban area. The survey conducted the face to face interview at household and individual level to gather the information. The survey was representative of 99 percent population of India living in 26 states, however survey did not cover the union territories. The NFHS-2 collected data from 89,199 women belongings to 91,196 households from the 438 districts of India. The overall response rate for nation was 96 percent, however the variation in response rate was observed in some states. The survey was weighted and normalize to adjust the nonresponse for each state. In NFHS survey, information about height and weight of all the children under age three years (since 1995) were gathered to determine the anthropometric measurement. After deleting flagged cases and missing information present study choose the WHO criteria to measure the childhood stunting, wasting and underweight in India.

The auxiliary information for the study were taken from the 2001 census of India. The chosen covariates were district level data on household size, availability of separate kitchen for cooking, availability of improved source of drinking water and cooking fuel. It has been noted that there may be some other indicators which may affect the nutritional status of children. But the present study did not consider these indicators in the study due to unavailability of data.

Using covariates from the 2001 Census to estimate nutritional indicators in the survey may raise the issue of comparability. However, using a combination of variables to derive composite indices minimizes these effects. This is because not all the indicators get changed over the short period of time.

Methodology

In this section present study illustrates the theoretical framework which is used to produced small area estimates of childhood stunting and their method of precision across the district of India by following the approaches mentioned elsewhere. Let N_d and n_d is the population and samples sizes in the district d (d=1, 2, 3 …. D), where D=438 is the number of district (has been considered as small area) in the population. In easier way, we defined N_d as the total number of children under age 3 years in d^{th} district recorded in census. And n_d is the number of sampled children under age 3 years in d^{th} district recorded in survey. Therefore, nd is the small sample of total children recorded by the survey and Nd refers the total number of children recorded in census in the same period. Let us suppose that Y_d is the value of response variable y (stunting) in the district d, i.e. total number of stunted children in district d^{th}. Here, the subscript d denotes the quantities belonging to district d. Further we used two more subscript 's' and 'r' to represent sample and non-sample population. In more clear way, it is Y_{sd} and Y_{rd}, where Y_{sd} is the sample stunted children in district d and Y_{rd} is the non-sample stunted children in district d. Thus, the response variable Y_{sd} follows a binomial distribution with parameter n_d and $\pi_{d.}$ Where π_d is the probability of being stunted in district d. Y_{sd} and Y_{rd} are assumed as the independent binomial variables with common success probability $\pi_{d.}$

$$Y_{sd} \sim \text{Bin}\ (n_d, \pi_d)$$

$$Y_{rd} \sim \text{Bin}\ (N_d\text{-}n_d, \pi_d)$$

Let x_d be the k vector of the covariates for the district d. The model linking this success probability with the covariates is the logistic liner mixed model of the form-

$$logit(\pi d) = \left\{\frac{\pi d}{1-\pi d}\right\} = \eta d = x'_d\beta + u_d \qquad \ldots\ldots\ldots\ldots (1)$$

d=1, 2, 3……., 438

Here, $\pi_d = \exp(\eta d)\{1+\exp(\eta d)\}^{-1}$

and β is the k vector of unknown fixed effects parameters.

$u_d \sim N(0, \phi)$ is the random effect that accounts for between district variability beyond that explained by the covariates included in the model. Here we observe that model (1) shows the area (here area is district) level proportions (direct estimates) from the survey to the area level (district) level covariates. Oftenly, this type of model is called as "area-level" model in SAE terminology. The concept of area level model was firstly proposed by Fay and Herriot to predict the mean per capita income (PCI) in small geographical areas (less than 500 persons) within counties in the United States. Notably the Fay and Herriot method for SAE is based on the area level linear mixed model and their approach is applicable to a continuous variable. However, in contrast model (1) is the special case of a generalized linear mixed model (GLMM)

with logit link function and suitable for binary outcome variable (Breslow and Clayton 1993). In case of our study nature of outcome variable is binomial, therefore GLMM approaches is suitable in this context. This model has been described by Saei Chambers (2003) in context of SAE and by the definition the means of Y_{sd} Y_{rd} given u_d under model (1) are-

$$E(Ysd/u_d) = n_d[exp(x'_d\beta + u_d)\{1 + exp(x'_d\beta + u_d)\}^{-1}] \qquad \ldots\ldots\ldots\ldots\ldots (2)$$

$$E(Yrd/u_d) = (N_d - n_d)[exp(x'_d\beta + u_d)\{1 + exp(x'_d\beta + u_d)\}^{-1}] \qquad \ldots\ldots\ldots\ldots (3)$$

Let T_d is the total number of stunted children in district d, then

$$T_d = y_{sd} + y_{rd} \quad \text{(d=1, 2....438)}$$

The first term y_{sd} is the sample count (i.e., direct estimates from survey) for the census window whereas the second term y_{rd} is the nonsample count that is unknown. Thus, an estimate $T_d^{\wedge}$ of the total number of stunted children in district *d*, which is obtained by replacing y_{rd} by its predicted value under model (1). That is-

$$T_d^{\wedge} = y_{sd} + y_{rd}^{\wedge} = y_{sd} + (N_d - n_d)[exp(x'_d\beta + u_d)\{1 + exp(x'_d\beta + u_d)\}^{-1}]$$

………….. (4)

$\hat{T_d}$ was estimated using only children within the census window to ensure consistency between N_d and n_d. The proportion (p_d) of stunted children in a district d is obtained as the total number of children within district. Thus, an estimate of p_d is-

$$\hat{p_d} = \frac{\hat{T_d}}{N_d} \qquad ………….. (5)$$

For the estimation of unknown parameters in (4) or (5), present study used Penalized Quasi-Likelihood (PQL) estimation of β and u= (u_1, u_2…. u_d) with restricted maximum likelihood (REML) estimation of ϕ as described in Saei and Chambers (2003) and Manteiga et al. (2007). In particular, we adopted the Saei and Chambers (2003) algorithm for the parameter estimation.

The mean squared error (MSE) estimates are computed to assess the reliability of estimates and also, to construct the confidence interval for the estimates. Following Saei and Chambers (2003) and Manteiga et al. (2007), the MSE estimate of small area predictor (4) is given by

$$mse(\hat{p_d}) = M_1(\hat{\emptyset}) + M_2(\hat{\emptyset}) + 2M_3(\hat{\emptyset}) \; ………….. (6)$$

In the equation (6) the first two components M_1 and M_2 constitute the largest part of the overall MSE estimates. These are the MSE of the best linear unbiased predictor-type estimator when ϕ is known (Rao, 2014). However, the third component M_3 have the variability due to estimates of ϕ. For the analytical expression of the components of MSE the diagonal matrices $\hat{V_{sd}}\ and\ \hat{V_{rd}}$ which are defined by the variances of the sample and non-sample part respectively are given below.

$$\hat{V_{sd}} = diag\left\{n_d \hat{p_d}\left(1 - \hat{p_d}\right)\right\}$$

$$\hat{V_{rd}} = diag\left\{(N_d - n_d)\, \hat{p_d}\left(1 - \hat{p_d}\right)\right\}$$

Let us suppose that, A= $\{diag\ (N^{-1}{}_d)\}\hat{V_{rd}}$

$$B = \{diag\ (N^{-1}{}_d)\}\hat{V_{rd}} X_r \text{-A} \hat{T_s}\ \hat{V_{sd}} X_s$$

$$\hat{T}_s = \left(\emptyset^{-1} I_D + \hat{V_{sd}}\right)^{-1}$$

Here, $X_s and\ \ X_r$ are the sample and non-sample population part of auxiliary information and I_D is the identity matrix of order D. We can rewrite this-

$$\hat{T}_{11} = \left\{X_s{}' \hat{V_{sd}} X_s - \quad X_s{}' \hat{V_{sd}}\, \hat{T}_s \hat{V_{sd}} X_s\ \right\}^{-1}$$

$$\hat{T}_{22} = \hat{T}_s + \hat{T}_s \hat{V_{sd}} X_s T_{11} X_s{}' \ \ \widehat{V'}_{sd}\, \hat{T}_s$$

With the help of these notations and with equation 6, it can be written that

$$M_1\left(\widehat{\emptyset}\right) = A\hat{T}_s A'$$

$$M_2\left(\widehat{\emptyset}\right) = BT_{11}B' \text{ and}$$

$$M_3\left(\widehat{\emptyset}\right) = trace(\widehat{\nabla}_i\, \hat{\Sigma}\ \widehat{\nabla}_J{}'\, v(\widehat{\emptyset})), where\ \hat{\Sigma} = V_{sd}^{\wedge} + \widehat{\emptyset}\, I_D V_{sd}^{\wedge} V_{sd}^{\wedge}{}'$$

Here $v\left(\widehat{\emptyset}\right)$ is the asymptotic covariance matrix of the estimates of variance components $\widehat{\emptyset}$, which can be evaluated as the inverse of the appropriate Fisher information matrix for $\widehat{\emptyset}$. This also depends upon whether we are using maximum likelihood or restricted maximum likelihood (REML) estimates for $\widehat{\emptyset}$. For REML estimates for $\widehat{\emptyset}$

$$v\left(\widehat{\emptyset}\right) = 2\left(\widehat{\emptyset}^{-2}(D - 2t_1) + \widehat{\emptyset}^{-4} t_{11}\right)^{-1}$$

Where $t_1 = \widehat{\emptyset}^{-1} trace(\hat{T}_{22})$ and $t_{11} = trace(\hat{T}_{22}\hat{T}_{22})$

Let us suppose that $\Delta = A\hat{T}_s$

And $\widehat{\nabla}_i = \frac{\partial(\Delta_i)}{\partial\emptyset/\emptyset=\widehat{\emptyset}} = \partial\widehat{(A_i T_s) /}\, \partial\emptyset | \varphi = \hat{\varphi}$

Where A_i is the ith row of the matrix A. Similar analysis has been done for underweight and wasted children. All the analysis has been conducted in STATA.

Results and Discussions

Usually after using GLMM method of estimation generally two type of diagnosis procedure has been used for diagnostic purpose. The first diagnostic is **1) model diagnostic**, which is used to verify the assumption of underlying model and the second diagnostic is **2) Diagnostics for small area estimates,** which is used to validate the reliability of model based small area estimates. Both of the methods have been described below.

Model Diagnostics

Purpose of using model diagnostic in present study is to verify the assumptions of model. In methodology section it has been described that the present study had used the logit link function of binomial family, therefore district level random effects should follow the normal distribution with mean zero and variance ɸ. If study follow the assumption of model, then district level residuals are expected to be randomly distributed and not significantly different from the regression line y=0. where under Model (1) the district level residuals are given $rd = nd - xd\beta$. The distribution of district level residual given in figure 1a, 2a and 3a shows that district level residual is randomly distributed and line of fit does not significantly different from line y=0. The q-q plots (figures 1b,2b and 3b) also confirm the normality assumptions of the data. Therefore, the diagnostic procedures related to model are fully satisfied is the study.

Figure 1a. Distribution of the district level residuals for childhood stunting.

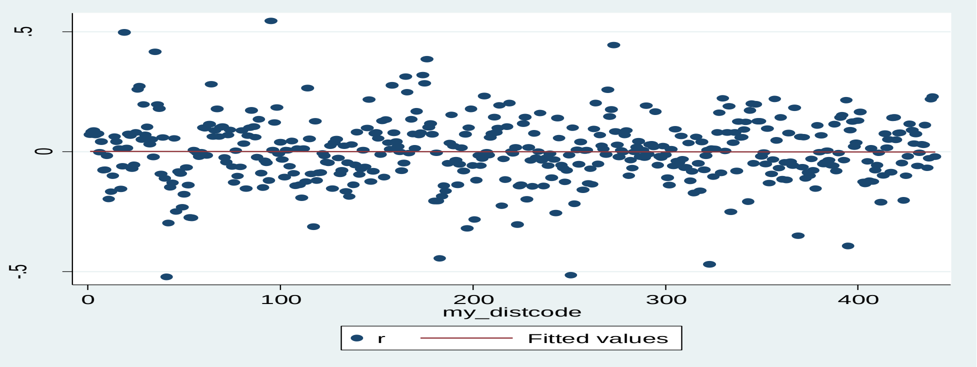


Figure 1b. **N**ormal q-q plot of the district level residuals for childhood stunting.

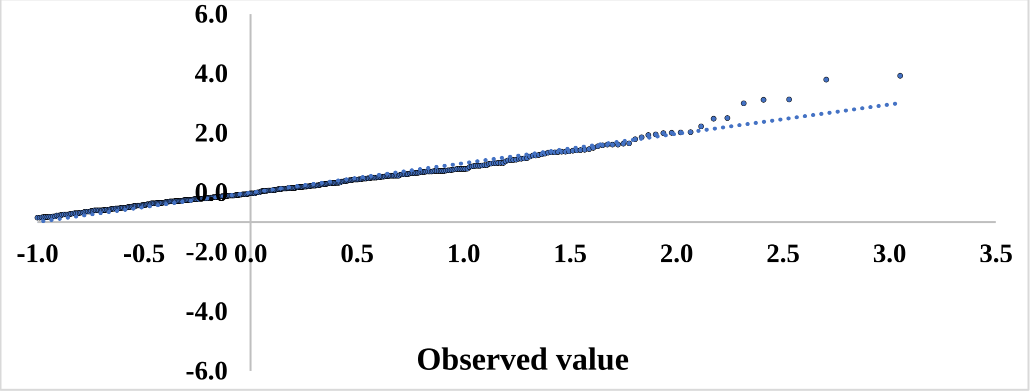


Figure 2a. Distribution of the district level residuals for childhood underweight.

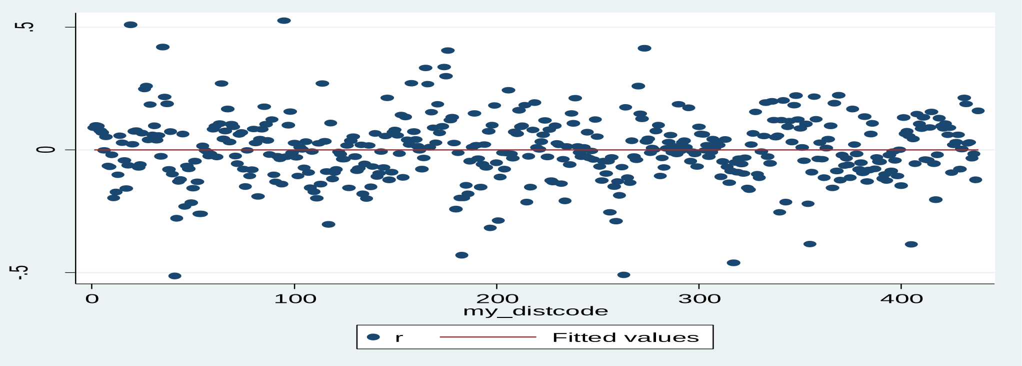


Figure 2b. Normal q-q plot of the district level residuals for childhood underweight.

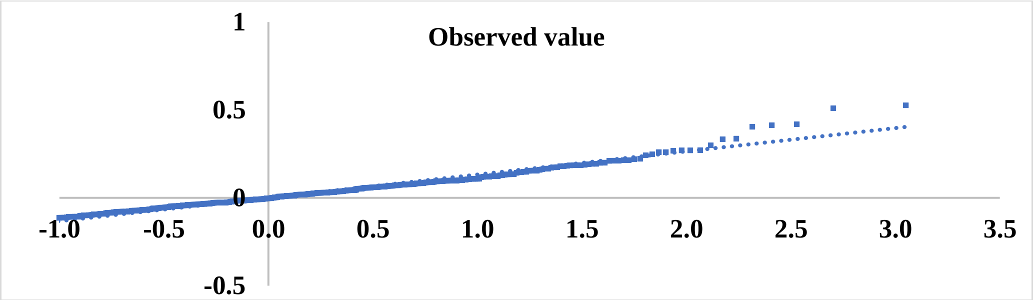


Figure 3a. Distribution of the district level residuals for childhood wasting.

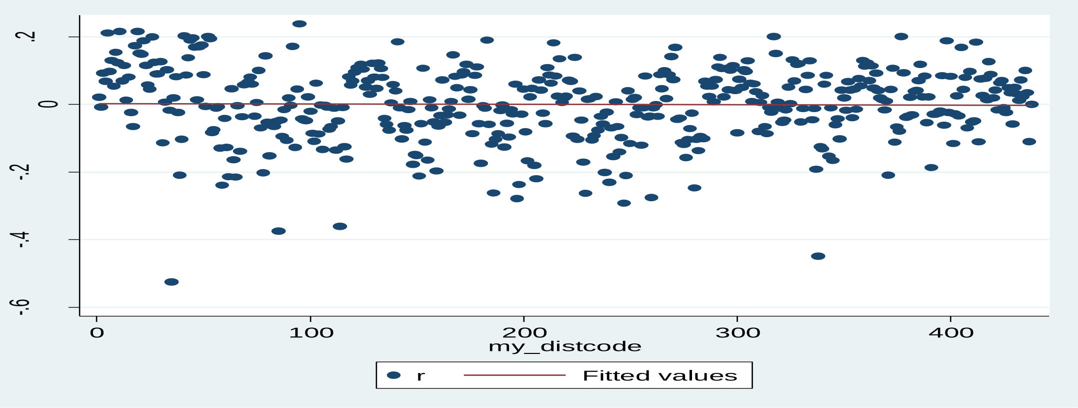
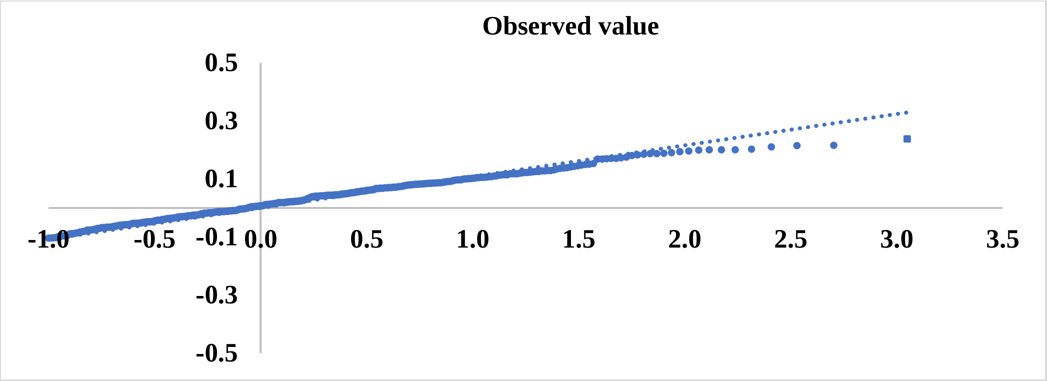


Figure 3b. Normal q-q plot of the district level residuals for childhood wasting.



Diagnostics for Small Area Estimates

The aim of this diagnostics procedure is to validate the reliability of the model-based small area estimates obtained by the GLMM. Present study used the bias diagnostics together with the coefficient of variation and computed the 95% CIs of the model-based estimates to validate the robustness of our model-based estimates relative to the direct estimates. The bias diagnostics are used to investigate if the model-based estimates are less extreme than the direct survey estimates. The direct estimates are calculated with survey weights. Figure 4,5 and 6 shows the bias scatter plot of the model-based estimates against that of the direct survey estimates. The figure 4 shows that the model-based estimates of stunting are less extreme than the direct survey estimates, and also reveals that the model-based estimates are shrinking towards the mean. Similar pattern of finding has been seen for underweight and wasted children. The study also highlighted that the districts having extreme direct estimates of either stunting, wasting or underweight were mainly due to small sample size.

We computed the coefficient of variation (CV) to assess the improved precision of the model-based estimates compared to the direct survey estimates. The CVs show the sampling variability as a percentage of the estimate. Estimates with large CVs are considered unreliable. There are no internationally accepted tables available that allow us to judge what is “too large.” Nonetheless the estimated CVs show that the model-based estimates have a higher degree of reliability than the (nonzero) direct survey estimates. For stunting and underweight, the estimated CVs for the model-based estimates range between 11.2% and 24%. The estimated CVs for direct survey estimates range between 16% and 93%. And 16% and 138% for underweight respectively. For wasting, the estimated CVs for the direct-survey estimates range between 11.2% and 43%, but the model-based estimates range between 3.1% to 3.9 %.

Approximate CIs for the direct estimates were calculated assuming that a simple random sample generated the weighted proportions. This ignores the effects of differential weighting and clustering within districts that would further inflate the true standard errors of the direct estimates.

Figure 4: Bias diagnostic for childhood stunting

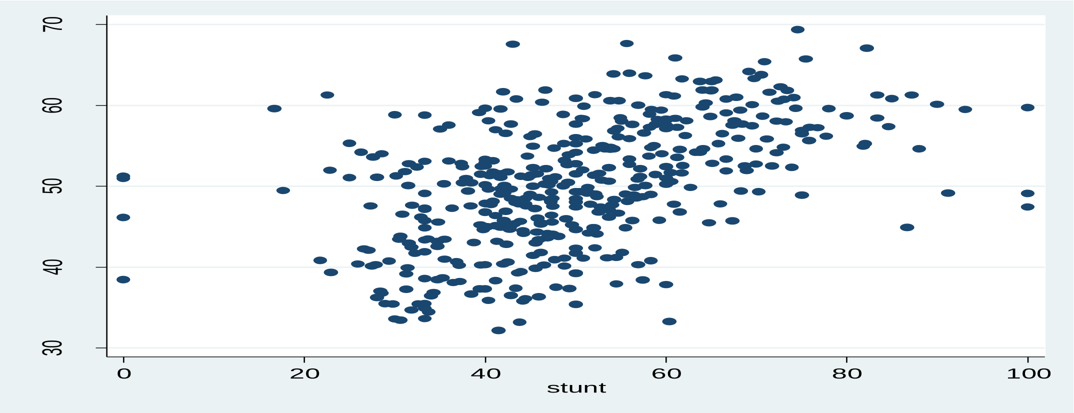


Figure 5: Bias diagnostic for childhood underweight

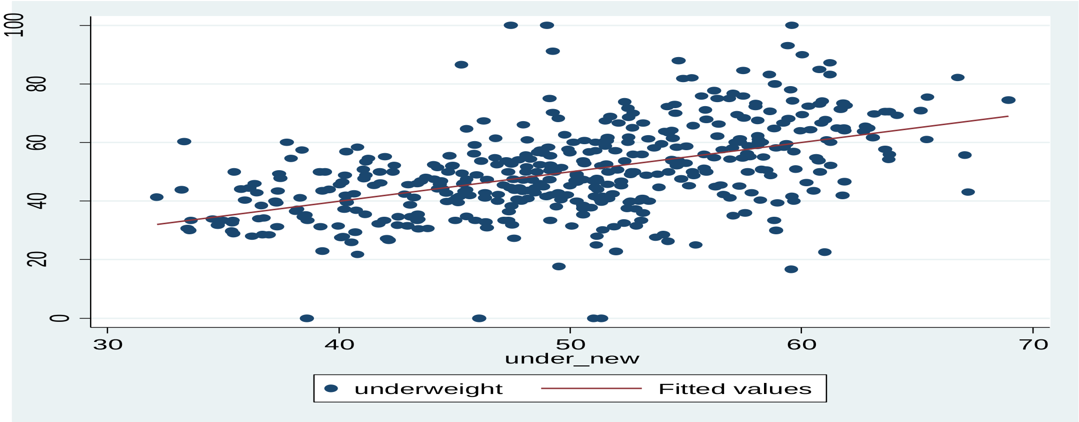
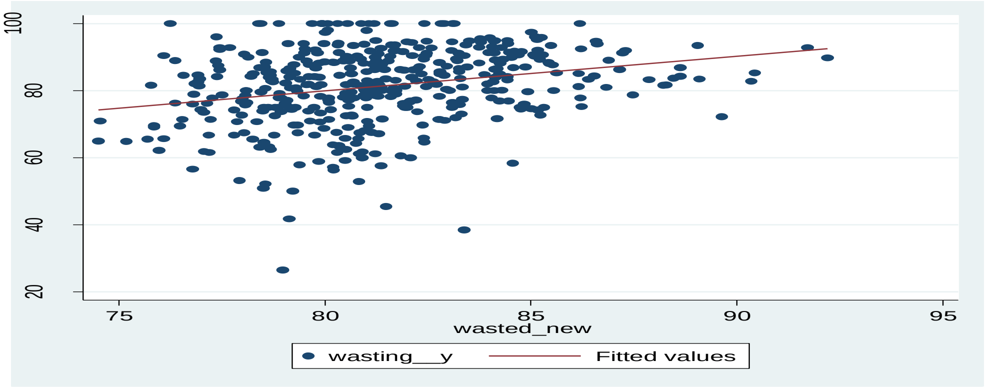


Figure 6: Bias diagnostic for childhood wasting



Discussion

The present study demonstrates the application of small area estimation techniques to derive district level statistics of stunting, wasting and underweight in India by using survey and census data. Although the SAE method for estimating proportions is well-developed (Saei and Chambers 2003; Manteiga et al. 2007), there is limited application in social sciences research. This article illustrates that the SAE method for estimating proportions is feasible with the type of outcome we have estimated.

An evaluation of the diagnostic measures confirms reasonably good precision of the model-based district estimates. The application of small area analysis is the first of its kind in the country, which lacks infrastructure and resources to collect representative data at the district level. The data from the census are usually limited as they tend to focus mainly on the basic socio-demographic and economic data. The IDHS, on the other hand, contributes to providing estimates at the regional and the national level. However, it is known that regional and national estimates usually mask variations (heterogeneity) at the district level and render little information for local level planning and allocation of resources.

In the case of India which has high levels of infant and child mortality, the availability of district-level data on health indicators is vital to monitoring health and facilitating a decentralized approach to health policy and planning. The district level estimates derived from the analysis also confirm a high degree of inequalities with regard to the uptake of institutional delivery care. The district-level variations seen in the distribution of malnutrition highlight the urgent need for appropriate policy interventions to monitor the supply and utilization of facilities in India. The targets set by the SDG to reduce hunger by zero seem a distant goal in most districts of India. This study has shown that with the availability of good auxiliary information and relevant survey data, policy-relevant local-level statistics could be derived to complement censuses which are limited in the amount of information they collect and are becoming less regular in India.

References

Rao, JNK. (2014). Small-Area Estimation. *Wiley StatsRef: Statistics Reference Online*, 1-8.

Saei, Ayoub, & Chambers, Ray. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects.

González-Manteiga, Wenceslao, Lombardía, María José, Molina, Isabel, Morales, Domingo, & Santamaría, Laureano. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational statistics & data analysis, 51*(5), 2720-2733.

Breslow, Norman E, & Clayton, David G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association, 88*(421), 9-25.