Determinants and geographical disparities of BMI in African Countries: a measurement error small area approach.

Determinanti e disparità geografiche del BMI nelle regioni africane: un modello di piccola area con errore di misurazione.

Serena Arima and Silvia Polettini

Abstract Food insecurity remains one of the greatest challenges in many African countries, hindering their economic development. Among related indicators, women's body mass index (BMI), measuring women's nutritional status, is a key indicator of the socio-economic development of a country. Despite recent intervention programmes, geographic and socio-economic disparities remain in the BMI distribution. Therefore, it would be important to rely on accurate estimates of women's mean BMI levels across domains. We consider a small area model with area-specific random effects that capture the regional differences in BMI levels. We propose a Bayesian model to investigate the role on BMI of a number of socio-economic characteristics such as age, wealth, parity, education, while accounting for regional variation. Since it is reasonable to assume that some of these variables are measured with error, we develop a suitable methodology and investigate the effect of neglecting measurement error in covariates on the assessment of the regression effects and on the prediction of area-specific BMI mean levels. We apply the proposed model to DHS data to explore the geographical variability of the BMI in two different regions, namely Ethiopia and Nigeria, and compare the determinants of women's nutritional status in these countries.

Abstract *L'insicurezza alimentare rimane una delle maggiori sfide per molti paesi dell'Africa. L'indice di massa corporea (BMI) femminile non fornisce solo una indicazione sullo stato di salute delle donne, ma è uno degli indicatori più utilizzati per valutare lo sviluppo culturale ed economico dei paesi. Nonostante le numerose azioni intraprese dai governi, ancora oggi permangono grandi disparità geografiche e socio-economiche a livello nutrizionale. Per la valutazione e la pianificazione delle politiche occorrono stime accurate dei livelli medi del BMI delle donne a livello regionale; a tale scopo, il lavoro propone un modello bayesiano di*

Serena Arima
Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza University of Rome, via del Castro Laurenziano 9, 00161 Roma, e-mail: serena.arima@uniroma1.it

Silvia Polettini
Name, Address of Institute e-mail: silvia.polettini@uniroma1.it

piccola area con effetti casuali atti a cogliere la variabilità intra-regionale della risposta. Il modello lega la variabile di interesse a una serie di caratteristiche socio-economiche individuali quali età, indice di benessere economico, livello di istruzione, numero di figli, tenendo conto della variabilità geografica del BMI. Tuttavia per alcune di tali variabili è ragionevole ipotizzare che siano state misurate con errore. Il modello proposto è stato quindi esteso includendo nella sua formulazione anche la presenza di errore di misurazione nelle covariate. Utilizzando i dati indiviuali risultanti dalle indagini DHS, il modello proposto viene applicato per stimare e confrontare i livelli medi del BMI delle donne di due Paesi, Etiopia e Nigeria, che presentano caratteristiche molto diverse. Le stime dei parametri del modello consentono inoltre di valutare e confrontare le determinanti dello stato nutrizionale delle donne nei due Paesi.

Key words: BMI, Food insecurity, measurement error, small area models

1 Introduction

Although the proportion and absolute number of chronically undernourished people has declined worldwide, progress has been uneven among developing countries. Women are clearly the most critical target group from a nutrition standpoint, therefore data on women's nutritional status is essential in monitoring the socioeconomic development of a country. Indeed it has a direct impact on women's health status, and several indirect effects through the multiple roles of women in generating income inside and outside the household, bearing children and being responsible for their families' nutrition and care. Not surprisingly, for many countries this aspect has been the object of prioritized interventions in the achievement of the Millennium Development Goals.

In this paper we study the Body Mass Index (BMI) in two African countries: Ethiopia and Nigeria. These two countries are very different from several points of view. First of all their geographical position: Nigeria is located in West Africa while Ethiopia is located in the Horn of Africa. Nigeria is often referred to as the "Giant of Africa", owing to its large population and economy. With 186 million inhabitants, Nigeria is the most populous country in Africa and the seventh most populous country in the world. With over 102 million inhabitants, Ethiopia is the most populous landlocked country in the world and the second most populous nation on the African continent.

Despite progress toward eliminating extreme poverty, Ethiopia remains one of the poorest countries in the world, due both to rapid population growth and a low starting base. More than 70% of Ethiopia's population is still employed in the agricultural sector, but services have surpassed agriculture as the principal source of GDP. According to the 2016 World Bank figures, life expectancy in Ethiopia is about 65 years while in Nigeria it is about 53. However, the human development index has been estimated to be equal to 0.448 in Ethiopia and 0.527 in Nigeria, where the

GDP per capita is more than three times as high as Ethiopia (5861$ vs 1734$, data in constant 2011 international dollars). Also, the literacy rate among the population aged 15 years and older is much lower in Ethiopia (39% in 2007) than in Nigeria (55% in the same year).

The figures published for Nigeria by the National Population Commission & ICF International in 2014 based on the 2013 DHS survey data indicate that the mean BMI among women aged 15-49 is 23.0 kg/m2. Instead, the 2011 Ethiopia DHS data give a mean BMI of 20.2 kg/m2. While rural/urban disparities exist in both countries, small geographical differences emerge from the figures published for Nigeria, although the North West has the lowest mean BMI (21.9 kg/m2). Invariably, mean BMI increases with increasing education level and shows a steady increase with increasing wealth. In Nigeria, 11% of women of reproductive age are thin or undernourished (BMI less than 18.5 kg/m2), as opposed to Ethiopia, where, according to the 2011 DHS figures, the same percentage amounts to 27%. Besides this, in Nigeria obesity is a public health problem, with a 17% of women being overweight (BMI of 25-29 kg/m2), and 8% obese (BMI of 30 kg/m2 or above). Notice that in Ethiopia only 6% of women are overweight or obese. In both countries the prevalence of overweight and obesity among women of reproductive age increases with age and is reportedly higher in urban areas than in rural areas. In addition, the wealth index seems to be strongly associated with being overweight or obese.

In this work we study the BMI of women aged 15-49: we consider data from the 2011 Ethiopia DHS and 2013 Nigeria DHS. Data are available at `www.measuredhs.com`. Both surveys were designed to provide population and health indicators at the national (urban and rural) and regional levels (for Ethiopia, the following 11 regions: Tigray, Affar, Amhara, Oromiya, Somali, Benishangul-Gumuz, SNNP, Gambela, Harari, and two city administrations, Addis Ababa and Dire-Dawa; for Nigeria, the 36 states plus the Federal Capital Territory are planned domains, but we consider for comparison the following 6 geo-political zones: South East, South South, South West, North Central, North West and North East). For both countries, the number of observations sampled in each region of interest is not particularly small. However, especially for Ethiopia, high geographical variability and the large population size make the problem of estimating the mean BMI at the domain level worth to be framed in a small area context. This also allows us to investigate the individual determinants of BMI while accounting for geographical variability.

We develop a small area model for studying the effect on BMI level of several potential explanatory variables: for each women, we have considered the number of sons, the education level, if they live in urban or rural centers, the age and the wealth index. The wealth index is built from from sample information on asset ownership, housing characteristics and water and sanitation facilities; it is obtained via a three-step procedure, based on principal components analysis, designed to take better account of urban-rural differences in wealth indicators. Being the result of a complex procedure, we treat the wealth index as a categorical covariate subject to misclassification. We also consider age, being self reported, as a continuous variable observed with error.

2 Small area models

In recent years, small area estimation has emerged as an important area of statistics as a tool for extracting the maximum information from sample survey data. Sample surveys are generally designed to provide estimates of totals and means of variables of interest for large subpopulations or domains. However, governments and policy makers are more and more interested in obtaining statistical summaries for smaller domains such as states or provinces. These domains are called small areas and are usually unplanned so that a small number of units is allocated in each of these areas. *Indirect estimators* are often employed in order to increase the effective domain sample size by borrowing strength from the related areas using linking models, census, administrative data and other auxiliary variables associated with the small areas. Depending on the type of data available, small area models are classified into two types: area-level and unit-level. Area level models relate the small area means to area-specific auxiliary variables. Such models are essential if unit level data are not available. Unit level models relate the unit values of the study variable to unit-specific auxiliary variables with known area means. In this paper we focus on unit-level models within a Bayesian framework. See [9] for an up-to-date review.
In this paper we focus on unit-level small area models given the availability of record-level data, described in Section 1.
Suppose there are m areas and let N_i be the known population size of area i. We denote by Y_{ij} the response of the $j-$th unit in the $i-$th area ($i = 1,...,m$; $j = 1,...,N_i$). A random sample of size n_i is drawn from the $i-$th area. The goal is to predict the small area means

$$\theta_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij} \tag{1}$$

based on the observed sample. To develop reliable estimates, auxiliary information, often in forms of covariates, measured at the unit or at the area level, may be exploited.Adopting a superpopulation approach to finite population sampling, a unit-level small area model is defined as

$$Y_{ij} = \alpha + \beta x_i + u_i + \varepsilon_{ij} \qquad i = 1,...,m; \quad j = 1,...,N_i \tag{2}$$

where x_i is an auxiliary variable observed for each area. ε_{ij} and u_i are assumed independent, $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0,\sigma_e^2)$ and $u_i \overset{\text{iid}}{\sim} N(0,\sigma_u^2)$. A random sample of size n_i is selected from the $i-$th small area ($i = 1,...,m$).
The model in (2) may be estimated based on maximum likelihood [2, 8], Empirical Bayes [5] and hierarchical Bayes approaches [3].
As stressed in [6, 7] auxiliary variables may be measured with error: It is well recognized that the presence of measurement error in covariates causes biases in estimated model parameters and leads to loss of power for detecting interesting relationships among variables. Several solutions exist, also in the small area literature: indeed corrections of the unit-level and area-level estimators have been proposed both in a frequentist and Bayesian context [10, 4, 1]. Relying on the model proposed in

[6], we extend it in order to account for measurement error in both continuous and discrete covariates and explore the impact of our procedure in the assessment of covariates' effect in a small area model designed to estimate regional mean BMI level in Nigeria and Ethiopia.

3 Measurement error small area models

Consider a finite population, whose units are divided into m small areas. As in the previous section, let the population size of the i-th area be N_i, $i = 1,\dots,m$. Let Y_{ij} be the value of the variable of interest associated with the j-th unit ($j = 1,\dots,N_i$) in the i-th area ($i = 1,\dots,m$). A random sample of size $n_i \geq 1$ is drawn from the $i-$th area population and the sample data are denoted by y_{ij} ($i = 1,\dots,m$; $j = 1,\dots,n_i$). For each area, we consider the following covariates: t_{ij} – the vector of p continuous or discrete covariates measured without error, w_{ij} and x_{ij} – respectively, a vector of q continuous covariates and h discrete variables (with a total of K categories), both measured with error. Denote by s_{ij} and z_{ij} the observed values of the latent w_{ij} and x_{ij}, respectively. We assume that the perturbation only depends on the unobserved category of the latent variable, so if $h > 1$ we assume independent misclassification. Without loss of generality, in what follows we assume $h = 1$.
Following the notation in [6], the proposed measurement error model can be written in the usual multi-stage way: for $j = 1,\dots,n_i$, $i = 1,\dots,m$ and for $k,k' = 1,\dots,K$

$$\text{Stage 1. } y_{ij} = \theta_{ij} + e_{ij} \qquad e_{ij} \overset{\text{iid}}{\sim} N(0,\sigma_e^2)$$

$$\text{Stage 2. } \theta_{ij} = t_{ij}^{'}\delta + w_{ij}^{'}\gamma + \sum_{k=1}^{K} I(x_{ij}=k)\beta_k + u_i \qquad u_i \overset{\text{iid}}{\sim} N(0,\sigma_u^2)$$

$$\text{Stage 3. } S_{ij}|w_{ij} \overset{\text{iid}}{\sim} N(w_{ij},\sigma_s^2), \qquad w_{ij} \overset{\text{iid}}{\sim} N(\mu_W,\Sigma_w)$$

$$\mu_W \sim N(0,\sigma_\mu^2 I)$$

$$Pr(Z_{ij}=k|X_{ij}=k') = p_{k'k} \qquad p_{k'\cdot} \sim \text{Dirichlet}(\alpha_{k',1},\dots,\alpha_{k',K})$$

$$Pr(X_{ij}=k') = \frac{1}{K}$$

We also assume that $\beta,\delta,\gamma,\sigma_e^2,\sigma_u^2,\sigma_s^2$ are, loosely speaking, a-priori mutually independent; in particular, $\beta \sim N(\mu_\beta,\sigma_\beta), \delta \sim N(\mu_\delta,\sigma_\delta), \gamma \sim N(\mu_\gamma,\sigma_\gamma)$, $\sigma_u^{-2} \sim Gamma(a_u,b_u)$, $\sigma_e^{-2} \sim Gamma(a_e,b_e)$, $\sigma_s^{-2} \sim Gamma(a_s,b_s)$.
Hyperparameters have been chosen to have flat priors. Finally, we assume $\Sigma_w = \sigma_w^2 I$, and σ_w^2, σ_μ^2 and $(\alpha_{k',1},\dots,\alpha_{k',K})$ all known.
Stage 3 describes the measurement error model for both continuous and discrete covariates: we assume that the continuous observable covariates S_{ij} are modeled as Gaussian variables centered at the true unobservable value w_{ij} with variability σ_s^2. The model for the unobservable continuous variables w_{ij} is assumed normal with

unknown mean and known variance.
Thanks to the multilevel model formulation, the measurement error mechanism need not to be assumed, which is a useful characteristic of our proposal. For the discrete covariates, the misclassification mechanism is specified according to the unknown $K \times K$ matrix P; we denote its $k'-$th row by $p_{k'.}$, whose entries, $p_{k'k}$, represent the probabilities $P(Z_{ij} = k|X_i j = k')$, $k = 1,\dots K$ that the observable variable Z_{ij} takes the $k-$th category, $k = 1\dots,K$when the true unobservable variable X_{ij} takes the $k'-$th category. We also assume that the misclassification probabilities are the same across subjects and that all the categories have the same probability $\frac{1}{K}$ to occur.
Over each $p_{k'.}$, $k' = 1,\dots,K$, we place a Dirichlet$(\alpha_{k',1},\dots,\alpha_{k',K})$ prior distribution, with known parameter. According to the above assumptions, we can estimate the transition matrix P jointly with all the other model parameters.
Using the Bayes theorem, the posterior distribution of the unknown parameter is proportional to the product of the likelihood and the prior distributions specified in Stage 4. As the posterior distribution cannot be derived analytically in closed form, we obtain samples from the posterior distribution using Gibbs sampling.

4 Results and comments

For each country we consider the estimates of the regression parameters obtained under the two models, with and without accounting for measurement error. Previous studies show that not accounting for measurement error may lead to inaccurate estimation of regression coefficients, which in turn may affect small area predictions. Figures 1 and 2 report the posterior distributions of the regression parameters under both models for Ethiopia and Nigeria, respectively. Under the measurement error model (top panels), the covariates' effects are all consistent with expectations. The BMI increases with the wealth index category: the poorest women are more likely to be underweight than the richest ones. Although expected, such an important effect of the wealth index has not been always confirmed in previous studies. Also, in both countries more educated women show a larger BMI than less educated ones, with an effect that increases with the educational level. The model also highlights the great disparity between urban and rural areas, where the women's undernutrition problem is more severe. The number of children ever born (parity) is found to affect women's nutritional status significantly only in Ethiopia, where the BMI decreases with parity. With respect to age, the model highlights positive linear association with BMI: younger women are more likely to be underweight than older ones, as well documented in the literature. On the other hand, under the model that ignores the measurement error in wealth index and age, the strong differential effect of the wealth index is lowered or, in the case of Ethiopia, disappears (see the bottom panel of Figure 1). This is also consistent with findings in the literature, that sporadically identifies this variable as important. With respect to the other parameters, while the meaning of the coefficients is coherent with those obtained with the proposed model, the variables' effects are considerably inflated. Noticeably, for the Nigeria data par-

ity is only significant under the unadjusted model. On the other hand, a linear effect of the wealth index on the BMI emerges quite clearly from the measurement error models in both countries, but not from the naive model. Moreover, the measurement error plays a different role in the two models: in the Ethiopia data, it strongly affects the parameters' estimates and, as expected, the posterior distribution of the P matrix is far from the diagonal one. On the other hand, in the Nigeria data the measurement error has a smaller impact on the estimates as the posterior distributions of the diagonal elements of P are concentrated around 0.9. In conclusion, the proposed model seems to be fairly robust with respect to misspecification of the measurement error mechanism.

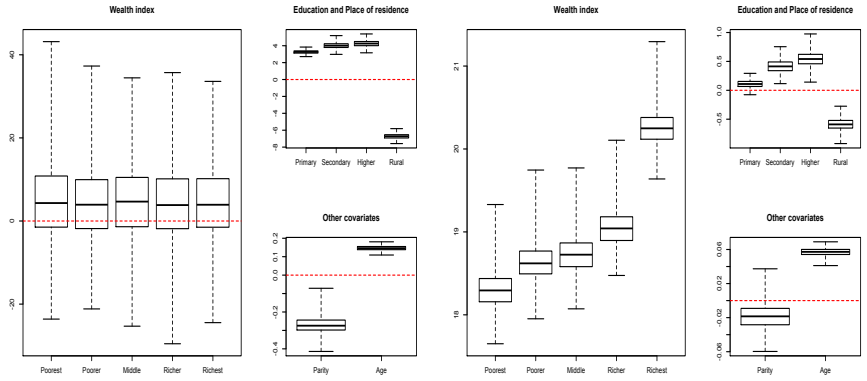

Fig. 1 Ethiopia data: posterior distributions of the model parameters under the proposed model (left panel) and under the assumption that the covariates are measured without error (right panel).

References

1. Arima, S., Datta, G.S. and Liseo, B. Bayesian Estimators for Small Area Models when Auxiliary Information is Measured with Error, Scandinavian Journal of Statistics, **42 (2)**,518–529, 2015
2. Battese, G.E., Harter, R.M. and Fuller, W.A. An error components model for prediction of county crop areas using survey and satellite data, Journal of the American Statistica Association, **83**, 28–36, 1988.
3. Datta, G.S. and Ghosh, M. Bayesian prediction in linear models: applications to small area estimation, Annals of Statistics, **19**, 1748–1770, 1991.
4. Datta, G.S., Rao, J.N.K. and Torabi, M. Pseudo-empirical Bayes estimation of small area means under a nested error linear regression model with functional measurement errors, Journal of Statistical Planning Inference, **140 (11)**, 2952–2962, 2010
5. Ghosh, M. and Rao, J.N.K. Small area estimation: an appraisal, Statistical Sciences, **9**, 55–93, 1994

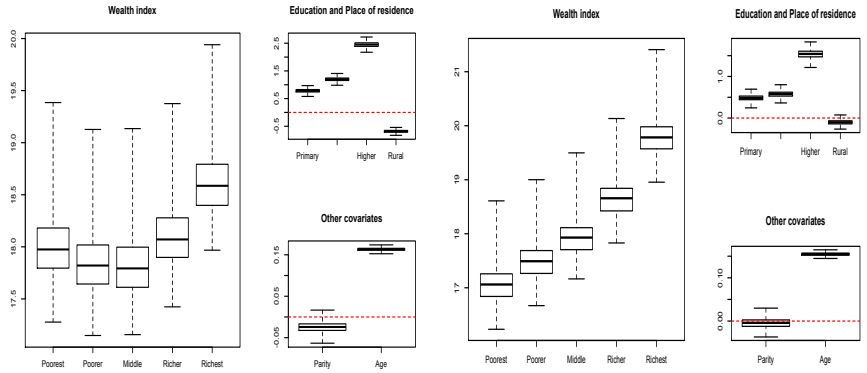

Fig. 2 Nigeria data: posterior distributions of the model parameters under the proposed model (left panel) and under the assumption that the covariates are measured without error (right panel).

6. Ghosh, M., Sinha, K. and Kim, D. Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models, Scandinavian Journal of Statistics, **33(3)**, 591-608, 2006
7. Ghosh, M. and Sinha, K. Empirical Bayes estimation in finite population sampling under functional measurement error models, Journal of Statistical Planning Inference **137**, 2759–2773, 2007
8. Prasad, N.G.N. and Rao, J.N.K. The estimation of mean squared error of small area estimators. Journal of the American Statistical Association, **85**, 163–171,1990.
9. Rao, J.N.K. and Molina, I.: Small Area Estimation, 2nd Edition, Wiley, Hoboken, New Jersey, 2015.
10. Ybarra, L.M.R. and Lohr, S.L. Small area estimation when auxiliary information is measured with error, Biometrika, **95(4)**, 919–931, 2008