

Journal of Computations & Modelling, vol.9, no.1, 2019, 1-9
ISSN: 1792-7625 (print), 1792-8850 (online)
Scienpress Ltd, 2019

Variance Estimation for Poverty Indicators via Linearization Techniques

Diogban Jakperik¹, Romanus Otieno Odhiambo² and George Otieno Orwa³

Abstract

In this study, an assessment of precision of poverty indicators is made with a view to improving its performance. A multiplicative bias reduction density function is used in estimating the poverty indicators and compared to the uniform, normal, and the nearest neighbor density estimators. Simulation results shows the practical potential of the multiplicative density estimator over its usual competitors especially when the sample size is large.

Mathematics Subject Classification: 62G07

Keywords: Variance estimation; poverty indicator; linearization technique; relative bias; density estimator

¹ Pan African University, Institute for Basic Sciences, Technology and Innovation.
Department of Statistics, Faculty of Mathematical Sciences, University for Development Studies,
Tamale, Navrongo Campus, Box 24, Ghana. E-mail: jdiogban@uds.edu.gh.

² Pan African University, Institute for Basic Sciences, Technology and Innovation.
Meru University of Science and Technology, P. O. Box 972-60200, Meru, Kenya.

³ Pan African University, Institute for Basic Sciences, Technology and Innovation.
Department of Statistics and Actuarial Science, Jomo Kenyatta University of Technology and
Agriculture, Nairobi, Kenya.

1 Introduction

Research efforts in the measurement of poverty indicators are recently being implemented using linearization techniques, which are seemingly preferred to the resampling methods because it is less labour intensive and time consuming [4] without sacrificing the gain in precision. Previously, linearization techniques were implemented especially for poverty and inequality indicators using the normal kernel density, this was shown to generate strong bias [5, 8]. [5] then proposed using the uniform and the k -nearest neighbor with logarithmic transformation to mitigate the bias. The reduction of the bias was still substantial after their methods were implemented. In this study, a fourth-order semiparametric density estimator is proposed, which significantly reduces the bias. This density estimator, reduces both bias and variance, or at worst preserves the variance of the ordinary kernel estimator and therefore makes it suitable for practical applications such as estimating poverty indicators.

2 Review of poverty indicators and their corresponding linearized variables

Suppose U is a population of size N distinct units u_1, \dots, u_N . For convenience, let u_k be represented by the index k . Let U be a population that has acceptable coverage of a target population. For every unit k , there is a corresponding measure y_k based on the characteristic of interest in the population. Because most survey data often contain tied observations due to rounding or range answers, making it difficult to sort into distinct objects for effective estimation of the density function, [5] increased these values randomly by small and negligible values uniformly to enable the data to be sorted. For a comprehensible review of these methods and applications, see [2, 3, 5, 10]. [5] gave the practical implementation of these methods. As poverty indicators are non-linear statistics, indeed they are rank statistics and therefore cannot be differentiated to the second order, the plausible approach to obtaining their linearized statistics is via generalized linearization [3, 5, 10]. Other methods to obtaining its variance are the resampling methods such as the Jackknife and the bootstrap. The generalized linearization technique uses the idea of *influence function* initially developed in the field of robust statistics by [6]. The definitions of

these poverty indicators considered in this study are simply stated below. Readers interested in details are referred to [5].

3 Poverty Indicators

3.1 Quantile

According to the fourth definition of [7], the quantile is defined as

$$Q_\alpha = y_{k-1} + (y_k - y_{k-1}) [\alpha N - (k - 1)] \quad (1)$$

where $\alpha N < k \leq \alpha N + 1$. The sample estimate of the quantile is

$$\hat{Q}_\alpha = y_{k-1} + (y_k - y_{k-1}) \left(\frac{\alpha \hat{N} - \hat{N}_{k-1}}{w_k} \right) \quad (2)$$

The linearized variable of an α -order quantile is given by

$$\hat{z}_k^{Q_\alpha} = -\frac{1}{f(\hat{Q}_\alpha)} \frac{1}{\hat{N}} \left[1_{[y_k \leq \hat{Q}_\alpha]} - \alpha \right] \quad (3)$$

[5].

The quantile estimates (2) and (3) will be used in estimating the poverty indicators which are discussed below.

3.2 Median income and at-risk-of-poverty threshold

Suppose $\hat{m} = \hat{Q}_{0.5}$ is the estimated median income of the sample. The At Risk of Poverty Threshold (*ARPT*) is defined as 60% of the median income:

$$ARPT = 0.6F^{-1}(0.5) \quad (4)$$

estimated by

$$\widehat{ARPT} = 0.6\hat{Q}_{0.5} = 0.6\hat{m}$$

This is an absolute measure that is scale-dependent. The linearized variable of the *ARPT* is proportional to that of the median income given by

$$\hat{z}_k^{ARPT} = I(ARPT)_k = 0.6I(MED)_k = -\frac{0.6}{f(\hat{m})} \frac{1}{\hat{N}} \left[1_{[y_k \leq \hat{m}]} - 0.5 \right] \quad (5)$$

[5].

3.3 At Risk of Poverty Rate

The At Risk of Poverty Rate ($ARPR$), where $ARPR \in [0, 1]$ defines the share of the population with an income below the $ARPT$: $ARPR = F(ARPT)$. It is also scale-dependent. The sample estimate is given by

$$\widehat{ARPR} = \frac{\sum_{y_k < \widehat{ARPT}} w_k}{\widehat{N}} \quad (6)$$

[5].

[12] defined the linearized variable of the $ARPR$ as

$$\begin{aligned} \hat{z}_k^{ARPR} &= \frac{1}{N} \left(1_{[y_k \leq \widehat{ARPT}]} - \widehat{ARPR} \right) - \frac{f(\widehat{ARPT})}{f(\hat{m})} \frac{0.6}{\widehat{N}} (1_{[y_k \leq \hat{m}]} - 0.5) \\ &= \frac{1}{\widehat{N}} \left(1_{[y_k \leq \widehat{ARPT}]} - \widehat{ARPR} \right) + f(\widehat{ARPT}) \hat{z}_k^{ARPT} \end{aligned} \quad (7)$$

3.4 Median income of individuals below the ARPT

The median income of individuals below the $ARPT$ is $m_p = F^{-1}(1/2F(ARPT))$. It is estimated in the same way like any other quantile but the exact definition may differ [5]. [10] defined the linearized variable of m_p in terms of the $ARPR$ as

$$\hat{z}_k^{m_p} = \frac{1}{f(\hat{m}_p)} \frac{\hat{z}_k^{ARPR}}{2} - \frac{1}{\widehat{N}} \left(1_{[y_k \leq \hat{m}_p]} - F(\hat{m}_p) \right) \quad (8)$$

3.5 Relative Median Poverty Gap

The relative median poverty gap ($RMPG$) is the relative difference between the $ARPT$ and the median income of individuals below the $ARPT$. If $RMPG = 0$, then the income of all "poor" individuals is equal to the $ARPT$, and $RMPG = 1$ if the income of all "poor" individuals is zero. It measures the extent to which "poor" individuals are poor;

$$RMPG = \frac{ARPT - m_p}{ARPT} \quad (9)$$

[5, 10]. The linearized variable of the *RMPG* as defined by [10] is

$$\hat{z}_k^{RMPG} = \frac{\hat{m}_p \hat{z}_k^{ARPT} - \widehat{ARPT} \hat{z}_k^{m_p}}{\widehat{ARPT}^2} \quad (10)$$

Here, the estimated income density function is involved four times: once in the estimation of \hat{z}_k^{ARPT} and three times in the estimation of $\hat{z}_k^{m_p}$.

3.6 Estimate of the income density function

Inferences on a finite population using a design-based study often rely on a design $P(S)$ to obtain representative sample of finite population U with size N . This way, the random variable of interest is the sample inclusion indicators whilst the rest are fixed. This makes the population income distribution a step function given by

$$F_y(x) = \sum_{k \in U} 1_{y_k \leq x} / N \quad (11)$$

which has no derivatives due to discontinuities. Suppose in superpopulation with a model-based study, the intent is not to justify the income density function, the distribution function can be smoothed artificially to become differentiable. Consequently, the function is not exactly a density function. Previously, researchers have smoothed the function using Gaussian kernel to obtain estimates of the income density function:

$$\begin{aligned} K(u) &= \frac{1}{h\sqrt{2\pi}} e^{-u^2/2}, \quad u = \frac{x - y_k}{h} \\ \hat{f}_1(x) &= \frac{1}{\hat{N}} \sum_{k \in S} w_k K\left(\frac{x - y_k}{h}\right) \\ &= \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k \in S} w_k \exp\left[-\frac{(x - y_k)^2}{2h^2}\right] \end{aligned} \quad (12)$$

with h as the bandwidth estimated by $\hat{h} = \hat{\sigma} \hat{N}^{-0.2}$; $\hat{\sigma}$ being estimate of the income standard deviation for the empirical income distribution given by $\hat{\sigma} = \sqrt{\frac{\sum_{k \in S} w_k y_k^2}{\hat{N}} - \bar{y}_w^2}$ [3, 5, 10]. It is worth noting that the presence of influential outliers as its often the case with income data can affect the estimate of the standard deviation. Therefore, [11] recommended the h for data that has positive skewness coefficient to be estimated by $h = 0.79 (\hat{Q}_{0.75} - \hat{Q}_{0.25}) \hat{N}^{-0.2}$. [10] observed that tied observations

as commonly found in survey data may affect the density function estimate and hence affect the precision of the linearization technique. In an attempt to address this defect, [5] made two propositions: First, they estimated the density at point x by estimating the density using the logarithm of x divided by its non-logarithmic value, valid for finite populations. The logarithm reduces the leveraging effect of the outliers present in the income data. Their estimator thus becomes

$$\hat{f}_2(x) = \hat{f}_v(v) = \frac{\hat{f}_v(v)}{x+a} = \frac{\hat{f}_y(\log(x+a))}{x+a} \quad (13)$$

where $a = (|\min_k(y_k)| + 1)$, a positive real number to cater for negative or zero incomes. Details can be found in [5]. Secondly, [5] estimated the density by using "nearest neighbor" with minimum bandwidth used by [?] for density estimation given by

$$\hat{f}_D(x, a, b) = \frac{1}{\hat{N}} \sum_{k \in S} K_D(y) \quad (14)$$

$$= \frac{1}{\hat{N}} \sum_{k \in S} w_k \frac{1}{b-a} 1_{y_k \in [a, b[} \quad (15)$$

$$= \frac{\hat{F}_Y(b) - \hat{F}_Y(a)}{b-a}, x \in [a, b[\quad (16)$$

with $\hat{F}_Y(x) = \sum_{k \in S} w_k 1_{y_k \leq x} / \hat{N}$ and $h = b - a$. Their method used at least p observations closer to point x based on minimum bandwidth $h(p) \geq h_{opt}$ defined by

$$h_{opt} = \frac{0.9 \min(\hat{\sigma}, \hat{Q}_{0.75} - \hat{Q}_{0.25})}{1.34 \sqrt[5]{\hat{N}}} \quad (17)$$

due [11] as rule of thumb for bandwidth determination. Consequently, they obtained the final density estimator using the proposed logarithmic approach as

$$\hat{f}_3(x) = \frac{\hat{f}(\log(x+a))}{x+a} \quad (18)$$

for details on the full implementation of this procedure, the reader should consult [5].

3.7 Multiplicative semiparametric bias reduction density estimator

In this section, a multiplicative semi-parametric biased reduction density estimator is proposed to effectively mitigate the challenge of bias in the estimation of

poverty indicators. The approach is to start with a parametric density estimate and multiply by a nonparametric kernel estimate. The general form of the density is

$$\begin{aligned}\hat{f}(x) &= f(x, \hat{\theta}) \hat{r}(x) \\ &= \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \frac{f(x, \hat{\theta})}{f(X_i, \hat{\theta})}\end{aligned}\quad (19)$$

where the nonparametric correction function is

$$\hat{r}(x) = \frac{1}{n} \sum_{i=1}^n \frac{K_h(X_i - x)}{f(X_i, \hat{\theta})}\quad (20)$$

Details and properties of the estimator can be found in Jakperik et al. (2018, unpublished).

4 Variance estimation

Let $\widehat{var}_{sim}(\hat{\theta})$ be a Monte Carlo variance estimate of a poverty indicator, $\hat{\theta}$. Suppose $\widehat{var}_{lin}(\hat{\theta})$ is the variance estimator based on linearization computed using the linearized variable, $\hat{z}_s^{\hat{\theta}}$ for every sample:

$$\widehat{var}_{lin}(\hat{\theta}) = \frac{N(N-n)}{n} var\left(\hat{z}_s^{\hat{\theta}}\right)\quad (21)$$

with n as the sample size for the simulations. Similarly,

$$var\left(\hat{z}_s^{\hat{\theta}}\right) = \frac{1}{n-1} \sum_{k \in S} \left(\hat{z}_{S,k}^{\hat{\theta}} - \bar{z}_S^{\hat{\theta}}\right)^2\quad (22)$$

where $\bar{z}_S^{\hat{\theta}} = n^{-1} \sum_s \hat{z}_{S,k}^{\hat{\theta}}$. This is strictly design-based and is obtained by substituting the linearized statistic for total into the relevant formula for variance according to the design used. In this study, the simple random sampling without replacement was used and hence its formulae above used in the relevant computations. The achieved reduction in bias using linearization is assessed by comparing the expected Monte Carlo value of the variance estimated using linearization, $E_{sim}[\widehat{var}_{lin}(\hat{\theta})]$ with the "true" Monte Carlo variance estimate, $\widehat{var}_{sim}(\hat{\theta})$, in terms of the relative bias:

$$RB[\widehat{var}_{lin}(\hat{\theta})] = \frac{E_{sim}[\widehat{var}_{lin}(\hat{\theta})] - \widehat{var}_{sim}(\hat{\theta})}{\widehat{var}_{sim}(\hat{\theta})}\quad (23)$$

[5].

5 Simulation results

Simulation studies performed to compare the performance of the proposed multiplicative density estimator to those used by [5] in their study. The results showed remarkable improvement in precision of the proposed multiplicative density function in estimating the poverty indicators considered in the study. One clear observation worthy of notice is the improvement in precision when the sample size increased from 500 to 1000. Therefore, it stands to reason that higher sample sizes may enhance the estimates even further. The table below presents the relative bias for the poverty indicators based on the different density functions.

Table 1: Evaluation of poverty indicators under different density functions

| Indicator | Sample size, $n = 500$ | | | | Sample size, $n = 1000$ | | | |
|-----------|------------------------|-------------|-------------|-------------|-------------------------|-------------|-------------|-------------|
| | \hat{f}_1 | \hat{f}_2 | \hat{f}_3 | \hat{f}_p | \hat{f}_1 | \hat{f}_2 | \hat{f}_3 | \hat{f}_p |
| ARPT | 0.06 | 0.04 | 0.12 | 0.03 | 0.05 | 0.02 | 0.09 | 0.01 |
| ARPR | 0.02 | 0.03 | 0.11 | 0.02 | -0.01 | 0.02 | 0.12 | 0.01 |
| RMPG | 0.42 | 0.22 | -0.13 | 0.09 | 0.40 | 0.20 | 0.10 | 0.04 |
| MEDP | 0.62 | -0.15 | 0.20 | 0.13 | 0.51 | -0.11 | 0.16 | 0.11 |
| MED | 0.08 | 0.06 | 0.14 | 0.04 | 0.04 | -0.03 | 0.08 | 0.02 |

Clearly, the estimators can either overestimate or underestimate the true parameters under consideration, resulting in the the relative bias taking on positive or negative values respectively depending on the nature of data used. The impressive issue about the results is that the multiplicative bias reduction density produces results with relative bias consistently lower than 5% which underscores its robustness.

References

- [1] Chauvet, G. and Goga, C., Linearization versus bootstrap for variance estimation of the change between gini indexes, *Survey Methodology*, (2018), 1742.
- [2] Demnati, A. and Rao, J.N.K., Linearization variance estimators for survey data, *Survey Methodology*, **30**, (2004), 17–34.

- [3] Deville, J.-C., Variance estimation for complex statistics and estimators: Linearization and residual techniques, *Survey methodology*, **25**(2), (1999), 219230.
- [4] Goga, C., Deville, J.-C., and Ruiz-Gazen, A., Use of functionals in linearization and composite estimation with applications to two-sample survey data, *Biometrika*, **96**(3), (2009), 691 709.
- [5] Graf, E. and Tille, Y., Variance estimation using linearization for poverty and social exclusion indicators, *Survey Methodology*, **40**(1), (2014), 6179.
- [6] Hampel, F.R., The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, **69**(346), (1974), 383393.
- [7] Hyndman, R.J. and Fan, Y., Sample quantiles in statistical packages, *The American Statistician*, **50**(4), (1996), 361365.
- [8] Karlis, D., Integrated systems of european social surveys - generalization of the existing framework to cover longitudinal and other complex aspects, *resreport*, (2016), EUROSTAT.
- [9] Mukhopadhyay, P., *Theory and Methods of Survey Sampling*, PHI Learning Private Limited, second edition, 2012.
- [10] Osier, G., Variance estimation for complex indicators of poverty and inequality using linearization techniques, *Survey Research Methods*, **3**(3), (2009), 167 195.
- [11] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.
- [12] Verma, V. and Betti, G., Taylor linearization sampling errors and design effects for poverty measures and other complex statistics, *Journal of Applied Statistics*, (2010).