Journaltitle of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: http://www.tandfonline.com/loi/gnst20

L_1Estimation of the median of a survey population

Gilbert W. Bassett Jr. & A.K. MD. E. Saleh

To cite this article: Gilbert W. Bassett Jr. & A.K. MD. E. Saleh (1994) L_1Estimation of the median of a survey population, Journaltitle of Nonparametric Statistics, 3:3-4, 277-283, DOI: 10.1080/10485259408832587

To link to this article: https://doi.org/10.1080/10485259408832587

Published online: 12 Apr 2007.

Nonparametric Statistics, Vol. 3, pp. 277–283
Reprints available directly from the publisher
Photocopying permitted by license only

L_1 ESTIMATION OF THE MEDIAN OF A SURVEY POPULATION

GILBERT W. BASSETT JR. and A. K. MD. E. SALEH

University of Illinois, Chicago and Carleton University

(Received October 5, 1993)

Improved estimates of a survey population parameter can be obtained by using information about a co-related auxiliary variable. The standard version of this problem concerns estimation of the population mean where ratio, difference, and regression methods for incorporating auxiliary information have been studied. Recently there has been interest in the related problem in which the median is the parameter of interest. Median versions of mean-based methods have been proposed and analyzed. Not yet considered however are methods based on the least absolute value (L_1) criterion for defining estimators. This approach is motivated by the criterion's close connection to the median. In this paper we introduce these methods and present properties in the context of the standard superpopulation model.

KEYWORDS: Median, L_1 estimation, survey, superpopulation model.

1. INTRODUCTION

There has been recent interest in estimating the median as well as the other quantiles of a survey population when there is information on a co-related auxiliary variable; see, Chambers and Dunstan (1986), Kuk (1988), Kuk and Mak (1989), and Rao *et al.* (1990). This is an extension of the standard problem of estimating the mean of a survey population. The question is how the standard methods are to be modified when the median rather than the mean is the parameter of interest.

Suppose we have n observations on a survey population that has N members. Associated with the survey variable, y, there is a co-related auxiliary variable, x, whose distribution is known. We want to use the sample data to estimate the population *median* of the y variable. An example might be estimation of median farm income in a population where we know the distribution of farm sizes – (say, acreage) – and where we have a random sample of $n \ll N$ farms that includes both income and acreage information.

Denote population values by $\{(y_j, x_j)\}, j = 1$ to N where the auxiliary values $x_j > 0$ are assumed known. To evaluate alternative estimation methods, let the existing survey population be modeled as the realization of a superpopulation; see Cassel *et al.* (1977), or Cochran (1977). That is, the population values themselves are presumed to be realizations of random variables, $Y_j = Y(x_j\beta)$, where

$$Y_j = x_j\beta + e_j v(x_j), \qquad j = 1, \ldots, N \tag{1.1}$$

β is unknown, $v(\cdot) > 0$ is a known function, and the e_j's are i.i.d. random variables with distribution F. The median of F is assumed to be zero so that the conditional *median* of $Y(x\beta)$ is $x\beta$. F_x denotes the known distribution of the auxiliary variable X. Except for specifying the median (instead of the mean) as the conditional location of $Y(x\beta)$, this formulation is identical to the standard superpopulation model. The two models would be identical if F were symmetric, but this restriction will not be imposed in the following.

The median of the survey population is a number, $y_m = M\{y_1, \ldots, y_N\}$ where $M\{\cdot\}$ denotes the median. This median is the realization of, $Y_m = M(Y(x_1\beta), \ldots, Y(x_N\beta))$, the median of the N random variables that generate the survey population. Let the sample values be denoted by (Y_i, x_i), $i = 1, \ldots, n$. It will be assumed that sample values are independent. The problem is how to use these sample values together with the known distribution of F_x to estimate the median y_m.

EXAMPLE. Consider the superpopulation model with $v(x) = 1$. Independent sample measurements are given in Table 1 and plotted in Figure 1 where x denotes farm acreage and y is farm income. A histogram of farm sizes for all farms is shown in Figure 2. It shows that the distribution of acreage in the population is skewed; the median is 500 and the mean is 600. The figure also shows the median acreage for the sampled data is $305 = \frac{1}{2}(289 + 332)$ thus indicating a predominance of small farms in the sample.

One possible estimate for median farm income is, $\$23{,}665 = \frac{1}{2}(20{,}455 + 26{,}876)$, the median income in the observed sample. This estimate however does not use the auxiliary information and hence does not reflect the predominance of small farms in the sample.

An estimate that does use sample information on x is based on a variation of the standard procedure for estimating the mean. The β parameter is estimated by generalized least squares (GLS), which with $v(x) = 1$ is the same as ordinary least squares (OLS). This β estimate can then be used to forecast Y at the median acreage of 500. In this case the estimate for median income is 500 times the OLS estimate, or $\$58{,}800 = 500 \times 117.6$. This method is identical to what is done in

Table 1. Sample data

Farm income	Farm size
10041	289
11876	244
20234	322
20455	145
268876	245
38444	455
445567	500
120333	550

Mean farm size in the population = 600
Median farm size in the population = 500

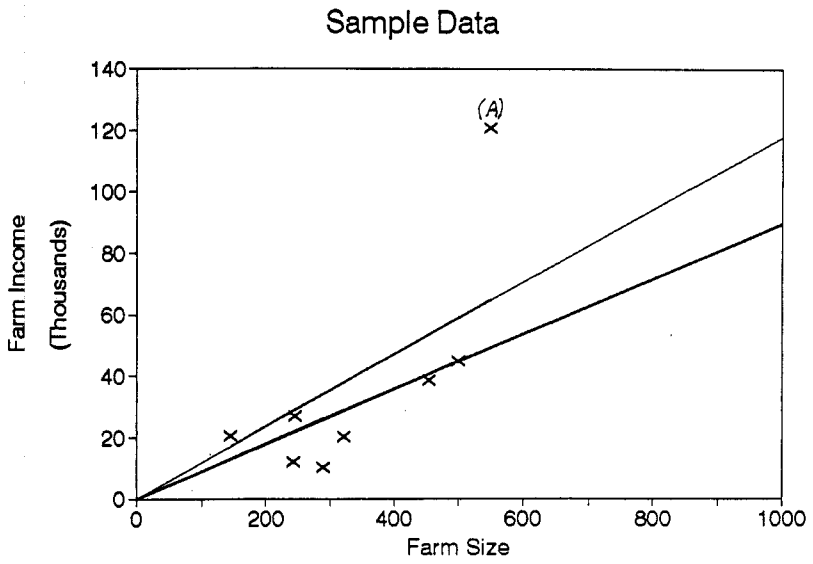

Figure 1

estimating the mean except that the median of the auxiliary variable is used in place of the mean. Notice that the two procedures would yield the same estimate if the mean and median values of the population auxiliary variable happened to be equal.

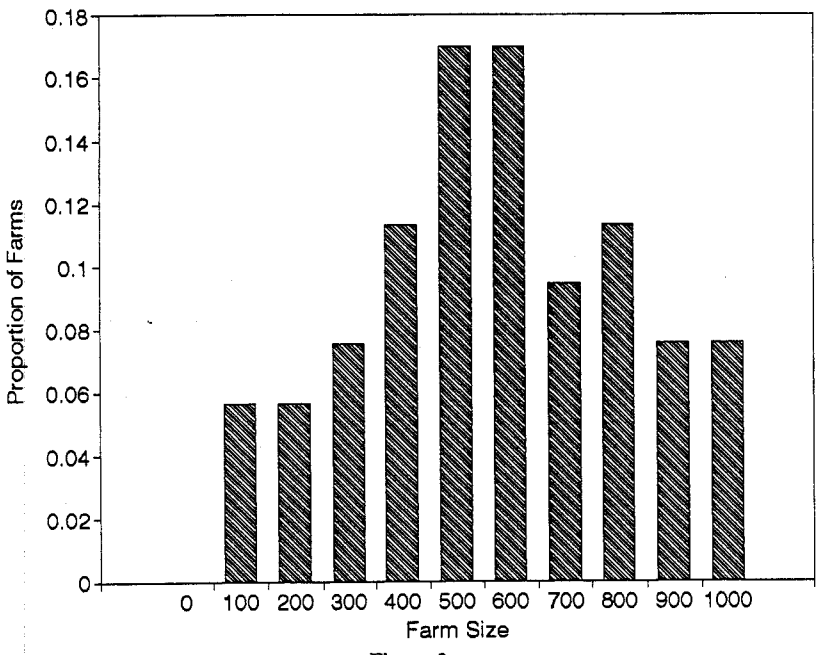

Figure 2

Our proposal for estimating median income uses the auxiliary information on x, but it is based on a least absolute value (L_1) estimate for β. The intuition that this will lead to better estimates comes from the well-known relation between the median and the L_1 criterion; the median makes the expected absolute deviation as small as possible. Further, many properties of the ordinary median (as well as the other quantiles) extend to the regression case in the form of the L_1 estimator (and the other regression quantiles); see, e.g., Bassett and Koenker (1978), and Koenker and Bassett (1978).

The L_1 estimate of 89.1 for the slope is indicated in the figure. It yields a median farm income estimate of $\$44{,}567 = 500 \times 89.1$, which is considerably different from what is obtained using the least squares approach.

One consequence of using L_1 is that it gives a different estimate for the median and mean, even when the mean and median values of x are identical. This follows from the fact that the least squares and L_1 estimates for β will not generally be equal. One reason why this is a desirable feature is that in the special case where all of the x values are identical our proposal reduces to using the sample median whereas the OLS-based method would produce the sample mean as an estimate for the population median.

Another desirable feature of the L_1-approach is that it better mimics properties of the median. For the clearest example, consider what happens when the observation labeled (A) in Figure 1 moves upwards. The OLS estimated line is pulled to (A) and the resulting estimate of median income based on OLS drifts up with (A) to infinity. The single large observation in the survey population has a large influence on the OLS estimate, but does not change the population median. This contrasts with what happens under our proposal because the L_1 estimate is unaffected by the outlying observation and hence the increase in (A) has no impact on the estimate, just as there is no impact if the observation (A) was already greater than median population income.

Details of the L_1 method are described in the next section. Included is the definition of the generalized L_1 estimates when $v(\cdot) \neq 1$, and the definition of the estimate that conditions on observed sample values. The third section establishes some properties relative to the superpopulation model. The estimates are shown to be median unbiased and consistent, and the generalized L_1 estimate is shown to minimize asymptotic variance within a class of median unbiased estimators.

2. DEFINITION OF ESTIMATORS

Estimators for β

We define a generic L_1 estimator for β that depends on a nonstochastic weighting vector $\mathbf{a} = (a_1, \ldots, a_n) \in R^n$ where $a_i \geq 0, i = 1, \ldots, n$. That is we define $b(a)$ as the minimum of $\rho(b:\mathbf{a})$ where $\rho(b:\mathbf{a}) = \sum a_i |Y_i - x_i b|$. As described below, the weights are used to define the generalized L_1 estimator that accounts for the heterscedasticity due to the $v(x_i)$ in (1.1).

The ordinary and generalized L_1 estimators for β. The ordinary L_1 estimator is defined by setting $a = (1, \ldots, 1)$ in the definition of $b(a)$. The generalized L_1

estimator is denoted by β_{GL1} and it is defined by setting $a_i = [v(x_i)]^{-1}$; that is the a_i weights are inversely proportional to the (known) scale of the y_i's.

Estimates for y_m. Let $\mathbf{x}^s$ and $\mathbf{x}^{ns}$ denote vectors that partition the auxiliary population into sampled and nonsampled values, respectively. (The components of $\mathbf{x}^s$ are the same as the sampled values except that replicated observations are omitted). To each component x_j in $\mathbf{x}^s$ there is an observed y_j value; these determine the vector $\mathbf{y}^s$ of sampled values. To each of the nonsampled x_j's we associate an estimate of its corresponding survey value, $\hat{y}_j$. This determines the vector $\hat{\mathbf{y}}^{ns}$ and then $(\mathbf{y}^s \& \hat{\mathbf{y}}^{ns})$, where the latter is an estimate of the entire survey population.

The estimate for the median is given by $M(\mathbf{y}^s \& \hat{\mathbf{y}}^{ns})$ where M is the median of the indicated set of values. This expresses the estimate in terms of the observed sample values and the forecasted values for the nonsampled members of the population.

When n is much smaller than N the estimate of the median is almost entirely determined by the forecasted values for the nonsampled part of the population; that is, the median $M(\mathbf{y} \& \hat{\mathbf{y}}^{ns})$ will be approximately the same as $M(\hat{\mathbf{y}}^{ns})$. In the case where there is no conditioning on the observed data we are essentially forecasting the parameter, $\text{med}[Y(x^0\beta)]$, which is the conditional median of Y at x^0, the median value of the x_i's.

REMARKS. 1. It is sometimes useful to express $b(a)$ in terms of the directional derivatives of ρ, $b(\mathbf{a}) = \{b \mid \psi_-(b:a) \le 0 \le \psi_+(b:a)\}$ where,

$$\psi_+(b:a) = -\sum a_i \operatorname{sgn}^*(Y_i - x_i b; -x)x_i, \qquad \psi_-(b:a) = -\sum a_i \operatorname{sgn}^*(Y_i - x_i b; x_i)x_i.$$

and $\operatorname{sgn}^*(v; w)$ is $\operatorname{sgn}(v)$ when $w \neq 0$ and $\operatorname{sgn}(w)$ otherwise.

2. The minimum set of ρ can be an interval. This situation can be ruled out by setting: $\sum a_i s_i x_i \neq 0$, for all $\mathbf{s} \in S$; where S is all vectors $\mathbf{s} = (s_1, \ldots, s_n)$ with $s_i = \pm 1$. This makes the minimum set of ρ a single point (and not a nondegenerate interval) for all $\mathbf{y} \in R^n$ so that $b(a)$ is then a well-defined estimator. This condition is not restrictive as it can always be achieved by slightly perturbing any one of the x_i's or a_i's.

3. The names "ordinary" and "generalized" follow the same, albeit confusing, conventions used for least squares. The GLS estimator is typically defined in terms of the *true* covariance matrix, though it is sometimes also taken to refer to any estimate of the form $(X'\Omega^{-1}X)X'\Omega^{-1}y$ – even when Ω is not the true covariance matrix. This standard convention is reversed for OLS, which is always defined as $(X'X)^{-1}X'y$ even when the covariance matrix is not proportional to an identity matrix.

3. PROPERTIES OF THE GENERALIZED L_1 ESTIMATOR

In this section we describe sampling properties of the estimator $b(a)$ relative to the model (1.1). We first give a representation of the sampling distribution. While the estimator has no closed-form expression in terms of the data, its distribution can be easily expressed in terms of sums of independent random variables. This

is used to show that $b(a)$ is median unbiased and then to derive the asymptotic distribution.

The distribution of $b(a)$. Consider the event, $[b(a) - \beta < \delta]$. This says that the convex function ρ attains its minimum to the left of $\beta + \delta$, or that the left derivative at δ is positive so that,

$$P[b(a) - \beta < \delta] = P\left[-\sum a_i \operatorname{sgn}^*(e_i - x_i\delta; x_i)x_i > 0\right]$$

$$= P\left[\sum -a_i S(d_i)x_i > 0\right]$$

where $S(d_i) = \operatorname{sgn}^*(e_i - x_i\delta; x_i)$ and $d_i = F(x_i\delta)$. The random variable $S(d_i)$ takes values -1 and $+1$ with probability $F(x_i\delta)$ and $1 - F(x_i\delta)$, respectively. The distribution of $b(a)$ can now be investigated via the behavior of the simple two-valued random variables, $S(d_i)$.

DEFINITION. An estimator β' for β is median-unbiased if

$$\Pr[\beta' < \beta] = \Pr[\beta' > \beta],$$

This says that it is equally likely for the esimator to be on either side of the parameter being estimated; Brown (1947) first considered median unbiasedness; its relation to the median estimator is discussed in Bassett (1991).

THEOREM 3.1. Each $b(a)$ is median-unbiased.

Proof. This follows directly from $S(\frac{1}{2}) = -S(\frac{1}{2})$ and the expression for the distribution, $P[b(a) - \beta < 0] = P[-\sum a_i \operatorname{sgn}^*(e_i; x_i)x_i > 0] = P[\sum -a_i S(\frac{1}{2}))x_i > 0] = \frac{1}{2}$.

The Asymptotic Distribution of $b(a)$. For the asymptotic results we consider the model (1.1) with $N \to \infty$ where

$$N^{-1}\sum \frac{a_i^2 x_i^2}{v_i^2} \to Q(a, x) > 0$$

$$N^{-1}\sum \frac{a_i x_i^2}{v_i^2} \to W(a, x) > 0$$

and F is assumed to have a derivative f, which is positive in a neighborhood of the median.

THEOREM 3.2. Under the above assumptions, $n^{1/2}[b_N(a) - \beta]$ converges in distribution to a normal random variable with mean zero and variance,

$$\frac{Q}{4f(0)^2 W^2}.$$

COROLLARY. The asymptotic variance of $b(a)$ is a minimum when $b(a)$ is the generalized L_1 estimator.

Proofs. We initially consider the case, $v(x_i) = 1$. The probability of the event, $[(n^{1/2}[b_n(a) - \beta] < \delta]$ is

$$P\left[\sum -a_i \operatorname{sgn}^*(e_i - N^{-0.5}x_i\delta; x_i)x_i > 0\right] = P\left[\sum -a_i S(d_i)x_i > 0\right] \qquad (3.1)$$

where $d_i = F(N^{-0.5}x_i\delta)$. Subtracting the mean, dividing by $N^{1/2}$, and multiplying by -1 gives

$$P\left[N^{-1/2}\left[\sum -a_i x_i[S(d_i) - E(S(d_i))]\right]\right] < N^{-1/2}\left[\sum -a_i x_i[E(S(d_i))]\right]$$

The expectation and variance of $S(d_i)$ are respectively,

$$E(S(d_i)) = -2[F(N^{-1/2}x_i\delta) - 0.5] = -2N^{-1/2}f(0)x_i\delta + N^{-1/2}x_i o_N(1)\delta$$
$$\text{Var}[S(d_i)] = 4[F(N^{-1/2}x_i\delta)(1 - F(N^{-1/2}x_i\delta)] = 1 + o_N(1).$$

Substituting for the right hand side of (3.1) gives,

$$N^{-1/2}\left[\sum -a_i x_i[E(S(d_i))\right] = 2f(0)\delta N^{-1}\sum -a_i x_i^2 = 2f(0)W(a, x)\delta + o_N(1)$$

while the left hand side converges to a normal random variable with mean zero and variance,

$$\text{Var}\left[N^{-1/2}\left[\sum -a_i x_i[S(d_i) - E(S(d_i))]\right]\right] = N^{-1}\sum (-a_i x_i)^2 = Q(a, x) + o_N(1)$$

Hence $n^{1/2}[b_n(a) - \beta]$ converges in distribution to a normal random variable with mean zero and variance $Q/[4f(0)^2W^2]$ as required.

To prove the corollary use Schwarz inequality, $[\sum (a_i x_i)x_i]^2 \leq \sum (a_i x_i)^2 \sum x_i^2$, so that

$$\frac{Q}{W^2} = \frac{\sum a_i^2 x_i^2}{[\sum a_i x_i^2]^2} \geq \frac{\sum a_i^2 x_i^2}{\sum a_i^2 x_i^2 \sum x_i^2} = \frac{1}{\sum x_i^2}$$

which shows that the variance is a minimum when all of the a_i's are equal to one.

To prove the result for $v(x) \neq 1$, divide both sides of (1.1) by $v(x_i)$ so that the transformed model has equal scale. Now use the $v(x) = 1$ result on the transformed model where the auxiliary variables are $x_i/v(x_i)$, and this completes the proof.

References

Bassett, G. W. and Koenker, R. W. (1978). The Asymptotic Theory of Least Absolute Error Regression, *Journal of the American Statistical Association*, **73,** 363, 618–622.

Bassett, G. W. (1992). "The Gauss Markov Theorem for the Median", *L_1-Statistical Analysis and Related Methods*, Y. Dodge (editor), Elsevier, 23–30.

Brown, G. W. (1947). On Small-Sample Estimation, *Annals of Mathematical Statistics*, **18,** 582–585.

Cassel, C., Sarndal, C. and Wretman, J. (1977). *Foundations of Inference in Survey Sampling.* Wiley.

Chambers and Dunstan (1986). Estimating Distribution Functions from Survey Data. *Biometrika*, **73,** 3, 597–604.

Cochran, W. G. (1977). *Sampling Techniques*, third edition, Wiley, New York.

Koenker, R. W. and Bassett, G. W. (1978). Regression Quantiles, *Econometrica*, **46,** 1, 33–50.

Kuk, A. Y. C. (1988). Estimation of Distribution Functions and Medians under Sampling with Unequal Probabilities. *Biometrika*, **75,** 1 97–103.

Kuk, A. Y. C. and Mak, T. K. (1989). Median Estimation in the Presence of Auxiliary Information. *JRSS(b)*, **51,** 2, 261–269.

Rao, J. N. K., Kovar and Mantel (1990). "On Estimating Distribution Functions and Quantiles From Data Using Auxiliary Information". *Biometrika*, **77,** 2, 365–75.