Rapporto n. 237

Median estimation with auxiliary variables

Rosita De Paola

Dicembre 2012

Dipartimento di Statistica e Metodi Quantitativi
Università degli Studi di Milano Bicocca
Via Bicocca degli Arcimboldi 8 - 20126 Milano - Italia
Tel +39/02/64483102/3 - Fax +39/2/64483105
Segreteria di redazione: Andrea Bertolini

1 Introduction

When skewed distributions such as consumptions and incomes are studied, the median is considered the more appropriate measure of location.

The literature with regard to the estimation of the median is less extensive than the studies regarding the mean. Moreover the estimation of the median usually does not consider the use of auxiliary variables.

In the present work the estimation of the median has been taken into consideration using different methods of analysis.

First of all the estimation of the median without auxiliary information is analyzed. Then the method which considers the median of the auxiliary variable is the ratio estimator. Then two methods based on the regression estimator are analyzed : the first one considers the regression based on the median regression, the second one is based on the minimum square method.

The methods are compared selecting all possible samples from nine different small populations.

Methodology

1.1 Median estimation without auxiliary information

The estimate of the median without the use of auxiliary variables is reduced to calculation of the median of the sample values.

With regard to the distribution of the sample median, Chu, in 1955, shows that if the parent population is Normal, then the distribution of the sample median tends "rapidly" to normality.

Let a continuous population be given with cumulative distribution function $F(x)$ and median ξ (assumed to exist uniquely). For a sample of size $2n+1$, let $\tilde{x}$ denote the sample median. The distribution of $\widetilde{X}$, under certain conditions, is known to be asymptotically Normal with mean ξ and variance $\sigma_n^2 = \frac{1}{[f(\xi)]^2(2n+1)}$, where $f(x) = F^{'}(x)$ is the probability density function.

Normal parent population. Suppose that a sample of size $2n+1$ is drawn from a Normal population with mean ξ and variance σ^2. The distribution of $\widetilde{X}$ is then asymptotically Normal with mean ξ and variance $\frac{\pi\sigma^2}{2(2n+1)}$. It has been shown that if, for $x > 0$

$$\phi(x) - \phi(-x) = a(x)\sqrt{1 - exp[-(2/\pi)x^2]} \qquad (1)$$

where $a(x)$ is a function of $x > 0$.

[5] proved that $a(x) \leqq 1$ and tabulated $\frac{1}{a(x)-1}$ for a number of values of x ranging from 0.1 to 2 .

[3] gave several proofs of the same inequality and remarked that if $\sqrt{1-exp[-(2/\pi)x^2]}$ is used as an approximation to $\phi(x)-\phi(-x)$, " then the error committed is less than one per cent of the quantity approximated." $a(x)$>0.9929 for all $x > 0$.

For arbitrary $x > 0$ and $y > 0$, let

$$x_n = \sqrt{\pi/2}x/\sqrt{2n+1} \quad , \quad y = \sqrt{\pi/2}y/\sqrt{2n+1} \ .$$

Applying (1) to the upper and lower bounds, [1] obtained

$$H(y) - H(-x) \geqq min\left\{a(x_n), a(y_n)\right\} * B_n\sqrt{1-\frac{1}{2n+2}} *$$

$$* \left[\phi\left(y\sqrt{\frac{2n+2}{2n+1}}\right) - \phi\left(-x\sqrt{\frac{2n+2}{2n+1}}\right)\right],$$

$$H(y) - H(-x) \leqq B_n\sqrt{1-\frac{1}{2n}}\left[\phi\left(y\sqrt{\frac{2n}{2n+1}}\right) - \phi\left(-x\sqrt{\frac{2n}{2n+1}}\right)\right],$$

where $B_n = \left(\frac{1}{2}\right)^{2n+1} C_n\sqrt{2\pi}/\sqrt{2n+1}$, $\phi(x)$ and $a(x)$ are defined by (1) and

$$\phi(t) = \int_0^t (1/\sqrt{2\pi})exp(-\frac{1}{2}x^2)dx.$$

For more details see the work of Chu (1955).

1.2 Median estimation using the ratio estimator

Given the population median of the auxiliary variable X , the ratio estimator is

$$\hat{Y}_R = \frac{\hat{M}e(Y) * Me(X)}{\hat{M}e(X)} \tag{1}$$

ie correcting the estimate of the median obtained from the sample with ratio between the median of X and its estimate.

If $\hat{M}e(X) = 0$ then $\hat{Y}_R = \hat{M}e(Y)$.

MSE of ratio estimator

Let

$$\frac{\hat{M}e_Y}{\hat{M}e_X} - \frac{Me_Y}{Me_X} = \frac{\hat{M}e_Y - \frac{Me_Y}{Me_X}\hat{M}e_X}{\hat{M}e_X}.$$

Note that

$$\frac{1}{\hat{Me}_X} = \frac{1}{Me_X + (\hat{Me}_X - Me_X)} = \frac{1}{Me_X} \frac{1}{1 + \frac{\hat{Me}_X - Me_X}{Me_X}}$$

we proceed with the Taylor series expansion, stopping at the first order

$$\simeq \frac{1}{Me_X} \left(1 - \frac{\hat{Me}_X - Me_X}{Me_X}\right)$$

then

$$\left(\frac{\hat{Me}_Y}{\hat{Me}_X} - \frac{Me_Y}{Me_X}\right) \cong \left(\hat{Me}_Y - \frac{Me_Y}{Me_X}\hat{Me}_X\right) * \frac{1}{Me_X}\left(1 - \frac{\hat{Me}_X - Me_X}{Me_x}\right) =$$

$$= \frac{1}{Me_X}\left(Me_Y - \frac{\hat{Me}_Y}{Me_X}\hat{Me}_X\right) - \frac{1}{Me_X}\left(\hat{Me}_Y - \frac{Me_Y}{Me_X}\hat{Me}_X\right)\left(\frac{\hat{Me}_X - Me_X}{Me_X}\right).$$

Passing to the squares

$$\left(\frac{\hat{Me}_Y}{\hat{Me}_X} - \frac{Me_Y}{Me_X}\right)^2 =$$

$$= \frac{1}{Me_X^2}\left[\left(\hat{Me}_Y - Me_Y\right) - \left(\frac{Me_Y}{Me_X}\hat{Me}_X - Me_Y\right)\right]^2 +$$

$$+ \frac{2}{Me_X^2}\left(\hat{Me}_Y - \frac{Me_Y}{Me_X}\hat{Me}_X\right)^2 \left(\frac{\hat{Me}_X - Me_X}{Me_X}\right)^2$$

we define the expected value

$$E\left(\frac{\hat{Me}_Y}{\hat{Me}_X} - \frac{Me_Y}{Me_X}\right)^2 \simeq \frac{1}{Me_X^2}[E(\hat{Me}_Y - Me_y)^2 +$$

$$+ \frac{Me_Y^2}{Me_X^2}E(\hat{Me}_X - Me_X)^2 - 2\frac{Me_Y}{Me_X}E(\hat{Me}_Y - Me_Y)(\hat{Me}_X - Me_X)] +$$

$$+ \frac{1}{Me_X^2}E\left[(\hat{Me}_Y - Me_Y)^2(\hat{Me}_X - Me_X)^2\right].$$

The mean square error of the ratio estimator for the median is approximately given by

$$MSE(\hat{Me}_Y^R) = MSE\left(Me_X\frac{\hat{Me}_Y}{\hat{Me}_X}\right) = Me_X^2 * MSE\left(\frac{\hat{Me}_Y}{\hat{Me}_X}\right) =$$

$$= E\left[\left(\hat{Me}_Y - Me_Y\right)^2\right] + \frac{Me_Y^2}{Me_X^2}E(\hat{Me}_X - Me_X)^2 - 2\frac{Me_Y}{Me_X}[Cov(\hat{Me}_Y, \hat{Me}_X) +$$
$$+(Me_Y - E(\hat{Me}_Y))(Me_X - E(\hat{Me}_X)]$$

ie the mean square error of the ratio estimator of the median can be written as a function of the mean square of the median of Y, the mean square error for the ratio of X multiplied by the ratio between the population medians and the square of Y and X, the covariance between the estimate of the medians of X and Y and the product of the distortions of X and Y, multiplied by the ratio between the true median of Y and X.

The above result is valid for n sufficiently large.

1.3 Median Regression

Suppose that there is only one explanatory variable ($p = 1$), (Radaelli, 2004)[4] the problem is reduced to determining the parameteres a and b of the line :

$$\hat{y} = a + bx \tag{2}$$

that, given N points (x_i,y_i), make the minimum sum of the absolute values S of the residuals r_i:

$$S = \sum_{i=1}^{N} \mid y_i - \hat{y}_i \mid = \sum_{i=1}^{N} \mid y_i - a - bx_i \mid. \tag{3}$$

The problem has been resolved geometrically by Boscovich and subsequently formalized by Laplace (1786) , in the case it requires that the straight line passes throught the point which has coordinates equal to the arithmetic means of X and Y. In the exposition which follows reference will be made to the work of Otto J. Karst (1958) [2], which illustrates the methodology for determining the parameters of the straight line to the smallest absolute values, distinguishing:

1. the restricted problem , where the desired line passes throught any designated point (x^*, y^*), not necessarily one of the given set of points ;
2. the unrestricted problem , in which there are no limitations to the straight line.

We consider, in this paper, the restricted problem.

Given a set of points $\{x_i, y_i\} i = 1, 2, ..., n$ find the equation of the line 2 , through any point (x^*,y^*), not necessarily one of the given set, such that

$$S = \sum_{i=1}^{N} \mid y_i - a - bx_i \mid$$
$$= \sum_{i=1}^{N} \mid y_i - bx_i - (y^* - bx^*) \mid \tag{4}$$
$$\sum_{i=1}^{N} \mid (y_i - y^*) - b(x_i - x^*) \mid .$$

We first translate the origin to the point (x_i, y_i) by the transformation

$$x_i^{'} = x_i - x^* \tag{5}$$

$$y_i^{'} = y_i - y^*$$

for $i = 1, 2, ..., N$.
Given a set of points $\{y_i, x_i\} i = 1, 2, ..., n$, find the equation of the line

$$y^{'} = bx \tag{6}$$

such that

$$S = \sum \mid y_i - y_i^{'} \mid \tag{7}$$

is minimum, where

$$\hat{y}^{'} \equiv bx_i^{'}. \tag{8}$$

S may be written

$$S = \sum_{i=1}^{n} \mid y_i^{'} - \hat{y}_i^{'} \mid = \sum_{i=1}^{n} \mid y_i^{'} - bx_i^{'} \mid . \tag{9}$$

Hence, the problem becomes finding b in (9) such that S will be a minimum for a known set of points $\{x_i, y_i\}$, which are related to the given set $\{X_i, Y_i\}$through the transformation (5).

The minimum of S can be determinated by the following procedure.

a) Rank the $y_i^{'}/x_i^{'}$ in ascending algebraic order.

b) To $-\sum \mid x_i^{'} \mid$ add successive values of $2 \mid \tilde{x}_i \mid$until change in sign at $i = r$ signals the minimum point of S, since it indicates a change in slope of the S curve from negative to positive.

c) This minimum lies directly above the point $(y_i^{'}/x_i^{'}, 0)$. Hence, $b_i = y_i^{'}/x_i^{'}$ is the value of b for which S is a minimum.

Since S is a minimum for b, the equation of the line of best fit is

$$\hat{y} = \left(\frac{y_i^{'}}{x_i^{'}} \right) x \tag{10}$$

in the transformed coordinates, or

$$y^{'} - y^* = \left(\frac{y_i^{'}}{x_i^{'}} \right) (x - x^*) \tag{11}$$

in the original coordinates.

MSE of median regression

The mean square error of median regression can be written, for a known value of b, as a function of the mean square error of the median of Y, the mean square error of the median of X multiplied by the square of b and the covariance of the estimated median of Y and X, in formula

$$E\left(\hat{Me}_{MR} - Me_Y\right)^2 = E\left(\hat{Me}_Y + b(Me_X - \hat{Me}_X) - Me_Y\right)^2 \qquad (12)$$

$$= E[(\hat{Me}_Y - Me_Y)^2 + b^2(\hat{Me}_X - Me_X)^2 - 2b(\hat{Me}_Y - Me_Y)(\hat{Me}_X - Me_X)]$$

$$= E(\hat{Me}_Y - Me_Y)^2 + b^2E(\hat{Me}_X - Me_X)^2 - 2bCov(\hat{Me}_Y, \hat{Me}_X).$$

1.4 Estimate of the median using the regression method.

In a similar way to the estimate of the mean, one can estimate the median.

Given the estimate of b and the population median of the auxiliary variable X , the estimate of the median of Y using the regression method is :

$$Me(\hat{Y}_{lr}) = \hat{Me}(Y) + b(Me(X) - \hat{Me}(X)).$$

2 Applications

We proceeded by constructing the cases in which the auxiliary variable X is symmetric, positive asymmetric and negative asymmetric. A similar procedure was used for the cases of Y. We applied all methods to each sample and we evaluated the expexted value, the variance and the mean square error, in order to choose the best estimator.

2.1 Examples of real distribution of the estimators

In order to extract all the possible samples, let's suppose of knowing nine small populations realized combining the symmetric auxiliary variable X, the positive asymmetric variable X and the negative asymmetric variable X with the cases of symmetry, positive and negative asymmetry of the variable Y.

The different values are respectively:

in the first case, where the symmetric X has been used, the following six values are consider simmetric with respect to the median have been taken into consideration:

$$X < -(1, 2, 3, 4, 5, 6);$$

for the asymmetric positive X, the values are:

$$X < -(2, 2.3, 3.5, 4, 8, 12);$$

for the asymmetric negative X , the values are :

$$X < -(1, 1.5, 2.5, 9, 9.19.2).$$

As for the values of the variable Y are:
for the symmetric case :

$$Y < -(10, 20, 30, 40, 50, 60);$$

for the positive asymmetric case:

$$Y < -(20, 25, 35, 40, 80, 120)$$

and for the negative asymmetric case:

$$Y < -(10, 15, 25, 90, 91, 92).$$

The median populations on three cases is :

Y symmetric	Y positive asymmetric	Y negative asymmetric
35	37.5	57.5

Case 1: X symmetric
Calculation of the expected values, the variances and the mean square errors.

Expected values	**Y sym**	**Y pos asym**	**Y neg asym**
No auxiliary variable	35,115741	45.0434	55.8184
Median Regression	35	44.91662	55.6547
Ratio	35	43.82974	55.4128
Linear Regression	35	44.79012	55.64535

Table 3: Expected values : X symmetric

Table 3 shows that
-for **Y symmetric:** the estimators are unbiased;
-for **Y positive and negative asymmetric :** the estimators are biased.

To choose the best estimator, we observe the Tables 5 and 7 :
-for **Y symmetric :** we note that the estimators using different methods improve on the estimate of the median without the auxiliary variable, but the choice is indifferent between median regression, ratio estimator and linear regression;
-for **Y positive asymmetric :** the best estimator is obtained by the ratio estimator.
- for **Y negative asymmetric :** the best estimator is linear regression's estimator.

Variances	**Y sym**	**Y pos asym**	**Y neg asym**
No auxiliary variable	140.282422	569.2835	1236.7778
Median Regression	0	216.4359	352.5256
Ratio	0	101.3940	509.6775
Linear Regression	0	120.125	331.1316

Table 5: Variances : X symmetric

MSE	**Y sym**	**Y pos asym**	**Y neg asym**
No auxiliary variable	140.2958	626.1874	1239.606
Median Regression	0	271.4422	355.9308
Ratio	0	141.4597	559.9055
Linear Regression	0	173.2708	334.5713

Table 7: MSE : X symmetric

Case2: X positive asymmetric
Calculation of the expected values, the variances and the mean square errors.

Expected values	**Y sym**	**Y pos asym**	**Y neg asym**
No auxiliary variable	34.9614	45.0228	55.0653
Median Regression	32.7494	37.5	52.1591
Ratio	31.25634	37.5	49.4480
Linear Regression	31.7568	37.5	45.3866

Table 9: Expected values : X positive asymmetric

Table 9 shows that
- for **Y symmetric** and **Y negative asymmetric** : the estimators are biased, even if the estimator without the auxiliary variable provides a value closer to that of the population;
- for **Y positive asymmetric :** the estimators are unbiased.

Variances	**Y sym**	**Y pos asym**	**Y neg asym**
No auxiliary variable	139.5226	567.7029	1245.6588
Median Regression	39.7785	0	881.6117
Ratio	33.6375	0	649.72
Linear Regression	29.174	0	714.3191

Table 11: Variances : X positive asymmetric

MSE	Y sym	Y pos asym	Y neg asym
No auxiliary variable	139.5241	574.1473	1251.586
Median Regression	44.83862	0	910.024
Ratio	47.6525	0	796.4536
Linear Regression	39.6913	0	779.1537

Table 13: MSE : X positive asymmetric

Table 11 and 13 show that

- for **Y symmetric** and **Y positive asymmetric :** the best estimator is the linear regression estimator ;

-for **Y positive asymmetric :** the choice is indifferent between median regression, ratio estimator and linear regression.

Case 3 : X negative asymmetric

Calculation of the expected values, variances and mean square errors.

Expected values	Y sym	Y pos asym	Y neg asym
No auxiliary variable	34.9704	44.9955	54.9954
Median Regression	35.7282	45.7777	57.5
Ratio	49.7288	63.2085	57.5
Linear Regression	35.868	46.6717	57.5

Table 15: Expected values : X negative asymmetric

Table 15 shows that

- for **Y symmetric** and **Y positive asymmetric :** the estimators are biased. It can be seen that for **Y positive asymmetric**, the ratio estimator is more biased;

- for **Y negative asymmetric :** the estimatorr are unbiased.

Variances	Y sym	Y pos asym	Y neg asym
No auxiliary variable	137.5683	533.6495	1248.3118
Median Regression	33.9634	314.4212	0
Ratio	467.4734	803.6303	0
Linear Regression	48.6476	416.5605	0

Table 17: Variances : X negative asymmetric

MSE	**Y sym**	**Y pos asym**	**Y neg asym**
No auxiliary variable	137.5691	589.832	1254.585
Median Regression	34.4894	382.9416	0
Ratio	684.4112	1464.786	0
Linear Regression	49.401	500.6814	0

Table 19: MSE : X negative asymmetric

Tables 17 and 19 show that:

- for **Y symmetric** and **Y positive asymmetric** : the best estimator is the median regression' estimator. We observe a high value for ratio estimator, which worses the estimation with respect to the method without auxiliary variable. Probably the ratio estimator has a high MSE due to the fact that the intercept is high.

- for **Y negative asymmetric :** the choice about the best estimator is indifferent between median regression, ratio estimator and linear regression.

3 Conclusions

In this paper, we intended to analyze the way in which the auxiliary information could be profitably used in order to improve the accuracy in the median estimation. We have tried to get the most efficient estimator comparing the estimator of the median without auxiliary variable and some estimators which keep into account the knowledge of an auxiliary variable.

In the case of an auxiliary variable, we analyzed:

Ratio estimator, that in some cases, is the most efficient estimator. But when the intercept value is high, the mean square error is worse than the one of the other methods.

In a lot of cases the most efficient estimators are median regression and linear regression, even if it is quite difficult to establish an objective order between the two.

The method of the median regression improves almost in every cases if it is compared to the ratio method, when the choice of the best method of estimation of the median has been analyzed.

Comparing the different methods seems, that the choice of one method or another is not unique, but it depends on the case of study.

References

[1] J.T. Chu. On the distribution of the sample median. *The Annals of Mathematical Statistics*, 26(1):112–116, 1955.

[2] O.J. Karst. Linear curve fitting using least deviations. *Journal of the American Statistical Association*, pages 118–132, 1958.

[3] G. Pólya. Remarks on computing the probability integral in one and two dimensions. In *Proceeding of the first Berkeley symposium on mathematical statistics and probability*, pages 63–78, 1945.

[4] P. RADAELLI. *La Regressione Lineare con i Valori Assoluti.* PhD thesis, Italy, 2004.

[5] JD Williams. An approximation to the probability integral. *The Annals of Mathematical Statistics*, pages 363–365, 1946.