Università degli Studi di Milano-Bicocca

DOTTORATO DI RICERCA IN STATISTICA E APPLICAZIONI

Ciclo XXIV

MEDIAN ESTIMATION WITH AUXILIARY VARIABLES

Tesi di Dottorato di : Rosita De Paola

Relatore:
Chiar.mo Prof.
Angiola Pollastri

Anno Accademico 2011/2012

Median estimation with auxiliary variables

All'altra parte di me

Durante questi tre anni ho avuto il piacere di collaborare con diverse persone che hanno contribuito alla mia crescita umana e professionale, vorrei esprimere a tutti loro la mia gratitudine.

Un ringraziamento particolare va alla Prof.ssa Angiola Pollastri per essere stata sempre prensente e disponibile con i suoi suggerimenti e preziosi consigli, per avermi regalato i suoi sorrisi anche nelle giornate di studio più difficili.

Un vivo senso di riconoscimento va al Prof. Michele Zenga per i suoi magistrali insegnamenti e per la disponibilità che ha sempre mostrato nei miei confronti.

Un grazie colmo di sincero affetto va alla Prof.ssa Manuela Cazzaro per avermi sostenuta ed incoraggiata durante questo ultimo anno e per avermi dato la possibilità di fare esperianza di attività didattica.

Ringrazio i miei colleghi per la collabozione e per aver contribuito a rendere più liete le ore trascorse in ufficio, tra gioie e anche piccole delusioni : Erica, Federica, Ilaria, Lorenzo, Romuald, Edit.

In particolare, un sentito grazie a Francesco per essere stato un perfetto compagno di studi ed amico; ad Alberto per i suoi insegnamenti su R e non solo; ad Elvira per i suoi speciali interventi tecnici; ad Angela & Elvira, per avermi sempre supportata moralmente.

Grazie a Federica per la sua costante e confortante presenza durante tutto il mio percorso accademico.

Un GRAZIE va ai miei genitori, perchè oltre alla fiducia incondizionata per le scelte fatte, a loro devo il privilegio dello studio che mi hanno donato; grazie a mio fratello, perchè senza di lui una parte di me sarebbe vuota.

Grazie ad Eugenio, che con il suo amore, ha saputo comprendere il sacrificio del nostro tempo insieme per lo studio.

Infine, spero vivamente di aver lasciato una "parte" di me ai miei studenti e a coloro che hanno avuto o avranno la curiosità di leggere la tesi.

Grazie a tutti.

Contents

List of Tables

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 axiliary variables.

In the present study 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 of Kuk and Mak proposed in 1989 is exposed: this way of estimating the median is based on the knowledge of the population median of auxiliary variable X. Another 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.

Two experiments have been carried out in order to compare the methods proposed. First of all the methods are compared selecting all possible samples from nine different small populations.

The second application is based on the selection of couples of random numbers from a bivariate random variable distributed as a Bivariate Log-Normal distribution. Also in this situation the methods of estimation of the median are compared considering the expected values and mean square errors.

CHAPTER 1

Median of a character of a population

The median Me is an index of position and is the central value of the distribution when the data is sorted. Specifically:

given a set of N units ordered $x_{(1)}, x_{(2)}, ...x_{(N)}$ (ordered according to a character), the median is presented by the central modality, where the central unit means the collective unity that divides into two equal parts : one part consisting of the units having a modality lower than or equal to the central unit and one part formed by units that have a modality greater than or equal to the central unit.

To calculate the median, it is necessary for the variable to be quantitative or qualitative in non-decreasing ordered; the method of determination of the median varies according to the type of character and distribution.

*Example*1. Calculation of the median of the heights (in meters) of 5 students:

$$1.75 \quad 1.72 \quad 1.68 \quad 1.74 \quad 1.80.$$

First we order the stature of the boys in ascending order:

$$1.68 \quad 1.72 \quad 1.74 \quad 1.75 \quad 1.80.$$

For the population of size N=5, the median is given by $x_{(3)} = 1.74$.

Formally, we proceed as follows: since N is odd, the median is the value that occupies the central position of the sequence in a non-decreasing order:

$$Me = x_{(\frac{N+1}{2})},$$

in this example:

$$Me = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 1.74.$$

Example 2. Calculation of the median time of 6 athletes running 200 m. The times (in seconds) ordered for the 6 athletes are:

$$24.7 \quad 25.2 \quad 25.1 \quad 25.6 \quad 25.7 \quad 26.1.$$

First we order the time of the athletes in non-decreasing order:

$$24.7 \quad 25.1 \quad 25.2 \quad 25.6 \quad 25.7 \quad 26.1.$$

In this example, applying the formula:

$$Me = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\},$$

the median value can be read as the value between 3-rd and 4-th place, therefore approximated by:

$$Me = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{25.2 + 25.6}{2} \right\} = 25.4.$$

Then, if N is even, the median is taken as half the sum of the two middle values.

In frequency distributions for discrete values, the data are usually already sorted, we must then calculate the absolute cumulative frequencies, which are obtained by associating with each value the sum of their frequencies with all those that proceed it, and determine what the value is.

If the data are grouped into classes, the class, in which there is the median, is determined using the median cumulative absolute frequencies. To obtain the exact median value, a linear interpolation is applied between the two extremes of the class in which the median lies, assuming that the frequencies are distributed in a regular class.

Example 3. Calculation of the median in the case of frequency distributions.

The formula applied is as follows:

$$M = x_{(\frac{N+1}{2})} = l_j^- + \left(\frac{N+1}{2} - C_{j-1} - \frac{1}{2} \right) * \frac{a_j}{n_j}$$

where

- l_j^- is the lower end of the median class,
- $\frac{N+1}{2}$ is the position,
- C_{j-1} is the cumulative frequency of the class preceding the median class,
- a_j is the size of the median class
- n_j is the frequency of the median class.

We report the frequency distribution and cumulative frequencies of 130 students of a fotball school in Milan, classified according to height.

X	n_j	C_j	t

			from	to
160-165	10	10	1	10
165-170	25	35	11	35
170-180	45	80	36	80
180-190	35	115	81	115
190-195	13	128	116	128
195-200	2	130	129	130
Total	130			

First of all we obtain the position $t = \frac{N+1}{2} = \frac{130+1}{2} = 65.5$, this is between 38 and 80, which are positions 170-180 of the third class.

The third class represents the median class. Therefore, applying the formula we have

$$M = 170 + \left(\frac{130+1}{2} - 35 - \frac{1}{2} \right) * \frac{10}{45} = 176.66$$

this result allows us to state that “half” of the players has a height less than or equal to 176.66 and the other “half” has a height greater than or equal to 176.66.

Relationship between mean and median. Often the mean and the median have similar values. This happens especially when the distribution of the variable is symmetric. However, if the distribution has strong asymmetry, the two measures may differ significantly.

*Example*4. Consideration of the following points scored by three players in 11 matches:

Player a	1	1	1	2	2	2	2	2	3	3	3
Mean	2										
Median	2										
Player b	1	1	1	2	2	2	2	2	3	3	4
Mean	2.1										
Median	2										
Player c	1	1	1	2	2	2	2	2	3	3	14
Mean	3										
Median	2										

We observe that the median does not change between the three sets, while the mean changes slightly between players a and b and differs significantly for the last player. In this case the mean has been heavily influenced by the last recorded score. In this example, the median seems to be preferable to the mean if the aim is to summarize the distribution, as, it is more stable and not affected by the extreme values.

More formally we can say that the median has the property of robustness. This property is particularly useful when we suspect that some large and / or small values are very abnormal and the result of error detection or just highly unusual situations. The mean has the property of robustness because it is sensitive to the presence of extreme values.

So on the one hand we have that the arithmetic mean, calculated as a value only for quantitative traits, is more informative than the central value. On the other hand, we have that, since the arithmetic mean has the characteristic of being very informative, it has the mean disadvantage of being very sensitive to the presence of extreme values or outliers.

The median, however, could also be calculated only for ordered qualitative characteristics and is very insensitive to the presence of outliers.

Remark. The mean and the median calculated from the same data sets, as we have seen, may be different. This is not surprising because they correspond to two different definitions of the center of a distribution. When the mean and the median of the same distribution are very different, then it is advisable to bring them both.

Let us consider the following values of income:

$$\$\ 40000 \quad \$\ 50000 \quad \$\ 58000 \quad \$\ 60000 \quad \$\ 136000.$$

This is best summed from the median rather than from the arithmetic mean, because the median is not affected by the extreme value \$ 136000. In fact $Me = 58000\$$ and $\mu = 68800\$$.

Therefore, when you must decide whether it is to your advantage to accept a job, do not ask the person who is offering an average salary in the company. Before you choose to accept the offer, ask what the median salary is!

Properties of the median. The median has the following properties.

(1) **Internality**. The median is always between the minimum mode $x_{(1)}$ and maximum mode $x_{(N)}$ character.

(2) **Monotone (in the weak sense).** If we consider the two series 30,30,30 and 30,30,50 we can observe that the median in both the first and second case is 30 even though the second distribution is statistically greater.

(3) **Not necessarily associated.** To prove this statement is enough to show an example where the associative property is not applied. Consider the unit distribution $(2,2,2,3,3,4,4)$ and its partition into two partial distributions $(2,2,2,4)$ and $(3,3,4)$. Calculate the median of the two partial distributions: $Me_1 = 2$ and $Me_2 = 3$. We observe that given the values $(2,2,2,2,3,3,3)$ the median of the distribution is equal to 2, and is different if we have the following observations from that of initial distribution, equal to 3.

(4) **The median minimizes the sum of the absolute values of differences.**

$$\sum_{i=1}^{N} \mid x_i - A \mid \geq \sum_{i=1}^{N} \mid x_i - Me \mid$$

The sum of the absolute values of deviations from the median is less than or equal to the sum of the values of the deviations from any other value.

(5) **Robust measure.** In other words, the median is little affected by the extreme values of the distribution. For example, we can observe that the two distributions 1,2,3 and 1,2, 100 have the same median but the mean is different in the two situations. This property is particularly useful when we suspect that some large or small modes are very abnormal and the result of errors of detection or just highly unusual situations.

In terms of the estimators, it is known that the median is a more robust estimator of location parameter of a distribution than the arithmetic mean. To better understand the difference between the two estimators, consider a set of N values:

$$x_1, x_2, ...x_N.$$

The mean and median, are respectively:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$M = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2} & \text{if } N \text{ is even} \end{cases}$$

Let us increase one of the two values of a sufficiently large quantity enough $\delta > 0$

$$x_j^* = x_j + \delta.$$

The mean becomes:

$$\bar{x}^* = \bar{x} + \frac{\delta}{N}$$

therefore it varies considerably if δ is sufficiently high.

With regard to the situations can instead be present can be :

a) the median does not change if:

- the increased value was greater than the median of the original values;
- the increased value was less than the median of the original values and remained lower even after the increase.

b) if the increased value was less than the initial median and, as a result of the increase, has become greater than the initial median, the median becomes:

$$M = \begin{cases} x_{\left(\frac{N+1}{2}+1\right)} & \text{if } N \text{ is odd} \\ \frac{x_{\left(\frac{N}{2}+1\right)}+x_{\left(\frac{N}{2}+2\right)}}{2} & \text{if } N\text{is even} \end{cases}$$

observed that the median is more robust than the arithmetic mean.

1.1. Uses of the median

Whenever we wish to obtain information on the shape of a distribution, we can use graphical and non-graphical tools, such as comparing the median with first and third quartile or interquartile average, to achieve the concepts of symmetry or asymmetry and skewed distributions.

1.1.1. Bowley's measures of skewness.

Among skewness measures one of the earliest to be introduced was

$$b = \frac{(Q_3 - median) - (median - Q_1)}{Q_{3-}Q_1}$$

where $Q_1$and Q_3 are the first and third quartiles of a sample or a distribution [Groeneveld and Meeden, 1984].

This coefficient was introduced by A.L. Bowley in 1901, who wrote

> Skewness, relating to the shape,..., of a curve is appropriately measured by an absolute quantity, and we therefore need a ratio of two concrete measurements.

The coefficient b is the ratio of the difference of two non negative numbers $Q_3 - m$ and $m - Q_1$ divided by their sum, where m represents the sample or population median. Thus, one has $-1 \leq b \leq 1$, where values of b near 1 indicate strong right skewness of the sample or distribution ; in the case of a symmetric sample or distribution, clearly, $b = 0$ (Encyclopedia of Statistical Sciences update Vol. 2).

In 1896 Francis Galton, the famous geneticist and scientist, suggested the use of sample or population deciles to measures skewness (Encyclopedia of Statistical Sciences update Vol. 2) ; Galton recommended

$$g = \frac{D_8 - median}{median - D_2}$$

where $D_2$and D_8 are respectively the second and eighth deciles of a sample in this case. If D_2 and D_8 are replaced by $Q_1$and Q_3 respectively in the equation for g, then

$$g = \frac{1+b}{1-b}$$

a monotonic increasing function of b satisfyting $0 \leq g < \infty$, with $g = 1$ corresponding to symmetry.

In general, indicating the cumulative distribution function of a continuous random variable X with $F(x)$, the natural generalization of Bowley's coefficient of skewness is given by

$$\gamma_\alpha(F) = \frac{[F^{-1}(1-\alpha) - m_X] - [m_X - F^{-1}(\alpha)]}{F^{-1}(1-\alpha) - F^{-1}(\alpha)}$$

$0 < \alpha < \frac{1}{2}$,

where m_X is the median of X. This is a scale and location-free measure of the skewness of a distribution. If F is symmetric, $\gamma_F(\alpha) = 0$. This measure, which replaces $Q_3$and Q_1 in the definition of b by the upper and lower αth quantiles of a distribution, was suggested by F.N. David and N. L. Johnson. In analogy with b,one has

$$\mid \gamma_\alpha(F) \mid \leq 1, \qquad 0 < \alpha < \tfrac{1}{2},$$

with values of γ_α near 1 and -1 indicating extreme right and left skewness, respectively.

Hinckley used

$$\tau\left(\alpha, F\right) = \frac{1+\gamma_\alpha(F)}{1-\gamma_\alpha(F)} = \frac{F^{-1}(1-\alpha) - m_X}{m_X - F^{-1}(\alpha)},$$

$$0 < \alpha < \tfrac{1}{2},$$

as a "crude outlier-free measure of asymmetry" to investigate which powers of skewed random variables yield transformed random variables with distributions close to symmetry [Brentari, 1990].

1.1.2. Skewed distributions. A continuous random variable X is asymmetric if, setting a value $x < Me(X)$, and indicating with $P\left\{(x_1, x_2)\right\}$the probability that the r.v. X takes values in the open interval (x_1, x_2) , we have that the probabilities of the intervals $(x, Me(X))$and $(Me(X), 2Me(X) - x)$, with the same size, such that

$$P\left\{(x, Me(X)\right\} \neq P\left\{(Me(X), 2Me(X) - x)\right\}$$

or

$$\frac{1}{2} - \phi(x) \neq \phi(2Me(X) - x) - \frac{1}{2}.$$

It follows that :

$$\phi(x) + \phi(2Me(X) - x) \neq 1$$

for at least a finite range on the x axis with a probability measure different from 0 [Frosini, 1990].

When there is an inverse function of ϕ in the smallest interval $\mathbf{I}$ such that $P\left\{X \in I\right\} = 1$, a fully equivalent definition of asymmetry can be given by considering the intervals $\left(x^{'}, Me(X)\right)$ and $(Me(X), x^{''})$ such that

$$P\left\{(x', Me(X))\right\} = P\left\{\left(Me(X), x^{''}\right)\right\};$$

setting this probability equal to ν for $0 < \nu < 1/2$, therefore the equalities being valid

$$x^{'} = \phi^{-1}\left(\tfrac{\mathbf{I}}{2} - \nu\right) \quad ; \quad x^{''} = \phi^{-1}\left(\tfrac{\mathbf{I}}{2} + \nu\right) \quad ,$$

we can say that random variable X is asymmetric if the size of the ranges is not the same that is, if

$$Me(X) - x^{'} \neq x^{''} - Me(X)$$

or

$$\phi^{-1}\left(\frac{\mathbf{I}}{2} - \nu\right) + \phi^{-1}\left(\frac{\mathbf{I}}{2} + \nu\right) \neq 2Me(X)$$

for at least a finite range on the ν axis.

Definition 1. It defines a continuous distribution that is skewed to the left such that, setting a value $x < Me(X)$, the intervals $(x, Me(X))$ and $(Me(X), 2Me(X)-x)$, having the same size, satisfy the following inequality:

$$\frac{\mathbf{1}}{2} - \phi(x) \geq \phi(2Me(X) - x) - \frac{1}{2}$$

or

$$\phi(x) + \phi(2Me(X) - x) \leq 1 \quad -\infty < x < \infty \tag{1.1.1}$$

with strict inequality for at least a finite range on the x axis with a probability measure different from 0.

The meaning of 1.1.1 is obvious : the formula is valid if, for each interval symmetric with respect to the median, the left half has a probability measure not less than that of the right half.

In the case that the random variable has values greater than the median, and these are all values x such that $x_0 \leq x \leq x_1$, we have left if the skew 1.1.1 is satisfied by setting $Me(X) = (x_0 + x_1)/2$.

It defines a distribution that is skewed to the right if the inequality is

$$\phi(x) + \phi(2Me(X) - x) \geq \mathrm{I} \quad -\infty < x < \infty \tag{1.1.2}$$

with a strict inequality for at least a finite interval on the x axis with a probability measure different from 0.

Obviously, if 1.1.1 and 1.1.2 are simultaneously true, distribution is symmetrical with respect to the median and we have

$$\phi(x) + \phi(2Me(X) - x) = \mathrm{I} \quad -\infty < x < \infty \ . \tag{1.1.3}$$

The concept of skewness is a subspecies of asymmetry, it is worth remenbering that according to Boldrini:

> asymmetric unimodal distributions are not always ascending or discending.

He stated his opinion defining a skewed distribution: << when it resembles the Normal type, but the ascending branch is steeper or less steep than the descending branch, so that the maximum ordinate no longer separates two tendentially equal, but substantially different portions of the surface>>.

1.1.3. Mean, median, mode and skeweness. The parameters mean, median and mode are measures of the center of the distribution and so may be interpreted as location parameters. In a symmetric distribution, the mean and median coincide with a mode.

For an asymmetric distribution, comparison of measures of location may be used as measures of skewness; the first of these to be considered was the relationship between the mean, median and mode. About the turn of the century, Pearson proposed the following coefficient :

$$S_k = \frac{Mean - Mode}{S.d.}.$$

The sign of S_k gives the direction and its magnitude gives the extent of skewness.

If $S_k > 0$, the distribution is positively skewed, and if $S_k < 0$ it is negatively skewed. So far we have seen that S_k is strategically dependent upon mode. If mode is not defined for a distribution we cannot find S_k. But empirical relation between mean, median and mode states that , for a moderately symmetrical distribution, we have

$$Mode - Mean \approx 3(Median - Mean).$$

Hence Pearson's coefficient of skewness is defined in terms of median as

$$S_k = \frac{3(Mean - Median)}{S.d.}$$

In 1917 with limited success Doodson considered establishing this relationship for a class of densities with small Pearson skewness , and for the Pearson family. Haldane (1942) shows that for certain distributions cloe to normal in the sense of having small cumulants, Pearson's relationship more or less holds.

Although it has been recognized that Pearson's empirical relationship does not generally hold, it is still often stated that mean, mode and median occur either in this or in the reverse order. A sufficient condition for this to hold is that $1 - F(Median + x) - F(Median - x)$ are of one sign in $x > 0$, nonnegative, for example, giving $Mean > Median > Mode$ (for a unimodal distribution). A stronger condition for this to hold is that $f(Median + x) - f(Median - x)$ change sign once in $x > 0$, negative to positive, for example, giving $Mean > Median > Mode$ [Eisenhauer, 2002].

CHAPTER 2

Estimation of the median

2.1. Estimation of the Population Median

A parameter of practical interest is the median of a finite population. The population median is a value Me that divides the population approximately in half, so that approximately half of the population elements have values smaller than Me, whereas the other approximately half have values larger than Me.

We obtain the median of the population using data from a probability sample.

Let $y_1, y_2, ..., y_N$ be the values of the population elements for the study variable Y. For any given number $y(-\infty < y < \infty)$, the population distribution function $F(y)$ is defined as the proportion of elements in the population for which $y_k \leq y$. More formally, the stepwise increasing function $F(y)$ can be written as

$$F(y) = \frac{1}{N}(\#A_y)$$

where A_y is the set of population elements with y_k values not exceeding y, that is, A_y =$\{k : k \in U\}$and $\{y_k \leq y\}$and $\#A_y$ denotes the number of the elements in the set A_y.

The population median Me is now defined as

$$Me = F^{-1}(0.5) \tag{2.1.1}$$

where $F^{-1}(\cdot)$ is the inverse function on $F(\cdot)$.

To estimate the population median Me, using data y_k for $k \in s$, where s is a probability sample, Särndal et al. [2003] proposed a general procedure, which involves two steps:

(1) first, obtain an estimated distribution function, indicated by $\hat{F}(y)$.
(2) then, estimate $Me = F^{-1}(0.5)$ by

$$\hat{M}e = \hat{F}^{-1}(0.5) \tag{2.1.2}$$

where the inverse $\hat{F}^{-1}(\cdot)$ is to be understood in the same way as $F^{-1}(\cdot)$ above.

Remark 1. The technique in (1) and (2) can be summarized as follows: we first produce an estimated distribution function $\hat{F}(y)$ in the same way as we would have determined Me from $F(\cdot)$, if $F(\cdot)$ had been known. The technique can be generalized to estimation of any parameter θ defined as a function of the population distribution function $F(y)$, for example, any quantile of the population.

We now describe in closer detail the mechanics of estimating the median Me, including a confidence interval procedure due to Woodruff [1952].

A sample s is selected from the population by a sampling design $p(\cdot)$ with inclusion probabilities π_k, π_{kl}. The estimation of $F(y)$, for any given y, can be viewed as the already familiar problem of estimating a population mean. To see this, note that $F(y)$ can be expressed as a population mean,

$$F(y) = \frac{1}{N}\sum_U z_{k,y} = \frac{t_{z_y}}{N} = \bar{z}_{yU} \tag{2.1.3}$$

where $z_{k,y}$ is an indicator variable, defined for $k = 1, ..., N$, and for any given real number y, as

$$z_{k,y} = \begin{cases} 1 & \text{if } y_k \leq y \\ \\ 0 & \text{if } y_k > y \end{cases}$$

and $t_y = \sum_U y_k$.

Consequently let us estimate $F(y) = \bar{z}_{yU}$, for a given y, by an expression corresponding directly to the sample-weighted mean $\tilde{y}_s$ (it is defined as $\tilde{y}_s = \frac{\hat{t}_{y\pi}}{\hat{N}}$):

$$\hat{F}(y) = \tilde{z}_{y,s} = \frac{\hat{t}_{z_y,\pi}}{\hat{N}} = \frac{\sum_s \frac{z_{k,y}}{\pi_k}}{\sum_s \frac{1}{\pi_k}} = \frac{\sum_{s\cap A_y} \frac{1}{\pi_k}}{\sum_s \frac{1}{\pi_k}} \tag{2.1.4}$$

where $(s \cap A)$ is the set of sample elements with values $y_k \leq y$. Now, $\hat{F}_y$, the estimated distribution function, is a non-decreasing step function climbing from zero to one like $F(y)$. To divide by $\hat{N}$ in equation 2.1.3, rather than by N (if N is known) will often have advantages from a variance point of view; also to have $\hat{N}$ in the determination of 2.1.4 means that $\hat{F}(y)$ reaches the ultimate value of unit as y increases, which is a desiderable property.

The approximate variance of the estimator 2.1.4, as well as an estimator of this variance, can be obtained from:

$$\hat{V}(\tilde{y}_s) = \frac{1}{\hat{N}^2}\sum\sum_s \check{\Delta}_{kl}\left(\frac{y_k - \tilde{y}_s}{\pi_k}\right)\left(\frac{y_l - \tilde{y}_s}{\pi_l}\right)$$

where

$$\breve{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}.$$

$\Delta_{kl}= \pi_{kl} - \pi_k\pi_l$ is the covariance of I_k and I_l.

Now, $\hat{F}(y)$ becomes a tool for finding the desired estimator $\hat{Me}$ of the median Me. As prescribed by step (2) above, we set

$$\hat{Me} = \hat{F}^{-1}(0.5)$$

where $\hat{F}^{-1}$ is the inverse function of $\hat{F}(y)$ given by (3). Note that to find $\hat{Me}$ this way, we need not to compute $\hat{F}(y)$ for the whole range of $y-$values, but only the following center part of the population. As with 2.1.1, it may happen that 2.1.5 does not produce a unique value of $\hat{Me}$.

Remark 2. For computational purposes we can alternatively express the median estimator as follows. Denote the y_k-values of the sampled elements, arranged in increasing order of size, by

$$y_{1:s} \leq y_{2:s} \leq \cdots \leq y_{n_s:s}$$

and the corresponding inclusion probabilities π_k by

$$\pi_{1:s}, \pi_{2:s}, \ldots, \pi_{n_s:s}$$

.

Set $B_0 = 0$, and define the cumulative sums

$$B_1 = \frac{1}{\pi_{1:s}}$$

$$B_2 = \frac{1}{\pi_{1:s}} + \frac{1}{\pi_{2:s}}$$

and so on; in general, for $l = 1, ..., n_s$,

$$B_l = \sum_{j=1}^{l} \frac{1}{\pi_{j:s}}$$

Clearly, $B_{n_s} = \hat{N} = \sum_{j=1}^{n} {}^{1}/_{\pi_{j:s}} = \sum_s {}^{1}/_{\pi_k}$.

The median estimator can be written

$$\hat{Me} = \begin{cases} y_{l:s} & \text{if } B_{l-1} < 0.5\hat{N} < B_l \\ \\ \frac{1}{2}(y_{l:s} + y_{l+1:s}) & \text{if } B_l = 0.5\hat{N} \end{cases} \tag{2.1.5}$$

To calculate $\hat{Me}$, we first compute $\hat{N}_\pi$, and the examine then cumulative sums B_l until we find that for some l,

$$B_{l-1} < 0.5\hat{N} \text{ and } B_l > 0.5\hat{N}$$

in which case $\hat{Me} = y_{l:s}$, or we find that for some l,

$$B_l = 0.5\hat{N}$$

in simple random sampling $\hat{Me} = \frac{1}{2}(y_{l:s} + y_{l+1:s})$.

Example. A sample of size five is drawn from a of size population $N = 7$, the observed data are

y_k
10206
7603
8591
6284
9278
10614

Considering a sampling with replacement, the probability is equal to $P_i = \frac{1}{N}$ and the inclusion probability is $\pi_k = 1 - (1 - P_i)^n$; rearrangement in increasing order of size of y_k gives the following:

l	y_k	P_i	π_k	$1/\pi_k$
1	6284	0.143	0.487	2
2	7603	0.143	0.487	2
3	8591	0.143	0.487	2
4	9278	0.143	0.487	2
5	10206	0.143	0.487	2
6	10614	0.143	0.487	2
7	11569	0.143	0.487	2
$\sum$				14

We calculate firtst the estimated distribution function $\hat{F}$, then $\hat{Me}$ from 2.1.2. Since $\hat{N} = 14$, we have using equation 2.1.4, the following:

y	$\hat{F}(y)$
$y < 6284$	0
$6284 \leq y < 7603$	${}^{2}\!/_{14} = 0.143$
$7603 \leq y < 8591$	${}^{(2+2)}\!/_{14} = {}^{4}\!/_{14} = 0.29$
$8591 \leq y < 9278$	${}^{(2+2+2)}\!/_{14} = {}^{6}\!/_{14} = 0.43$
$9278 \leq y < 10206$	${}^{(2+2+2+2)}\!/_{14} = {}^{8}\!/_{14} = 0.57$
$10206 \leq y < 10614$	${}^{(2+2+2+2+2)}\!/_{14} = {}^{10}\!/_{14} = 0.71$
$10614 \leq y < 11569$	${}^{(2+2+2+2+2+2)}\!/_{14} = {}^{12}\!/_{14} = 0.85$
$11569 \leq y$	1

or, in graph form,

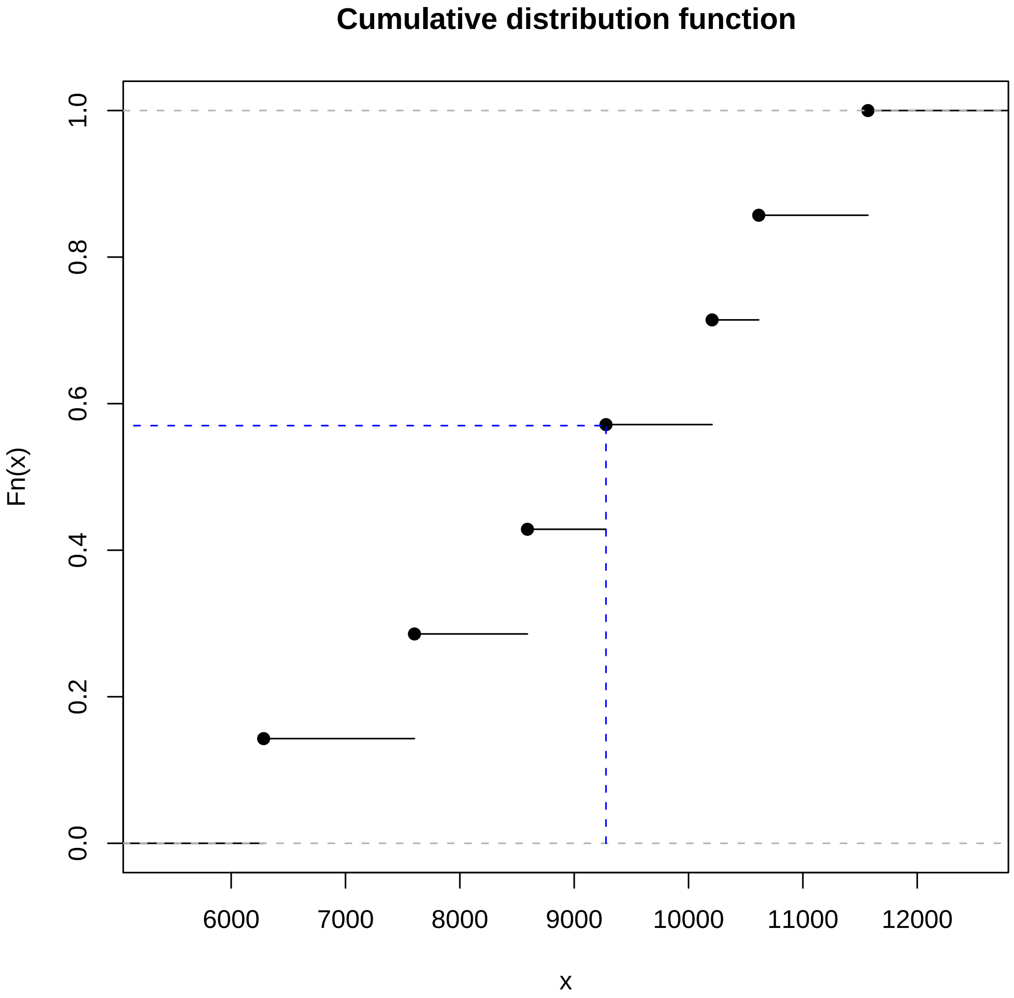

FIGURE 2.1.1. Cumulative distribution function

From the table, or from the graph, we find the estimated median to be

$$\hat{Me} = \hat{F}^{-1}(0.5) = 9278.$$

Alternatively, $\hat{Me}$ may be calculated from 2.1.5 as follows:

l	y_k	$1/\pi_k$	B_l
1	6284	2	2
2	7603	2	$2+2=4$
3	8591	2	$2+2+2=6$

4	9278	2	$2+2+2+2=8$
5	10206	2	$2+2+2+2+2=10$
6	10614	2	$2+2+2+2+2+2=12$
7	11569	2	$2+2+2+2+2+2+2=14$

Since $B_3 = 6 < 14/2 = \hat{N}/2$, and $B_4 = 8 > 14/2$, we obtain

$$\hat{M}e = y_4 = 9278.$$

2.1.1. Asymptotic distributions of estimators. Now we give the asymptotic distribution of the estimators $\hat{M}e_{YR}$, $\hat{M}e_{YP}$ and $\hat{M}e_{YS}$ as $N \to \infty$, $n \to \infty$ and $^n/_N \to f$, $0 \leq f < 1$. It is assumed that as $N \to \infty$ distribution of the bivariate variable (X, Y) approaches a continuous distribution with marginal densities $f_X(x)$ and $f_Y(y)$ for X and Y respectively. This assumption holds in particular under a super population model framework, treating the values of (X, Y) in the population as a realization of N independent observations from a continuous distribution; we assume also that $f_Y(Me_Y)$ and $f_X(Me_X)$ are positive.

Under these conditions, the sample median $\hat{M}e_Y$ is consistent and asymptotically normal [Gross, 1980] with expected value M_Y and variance

$$(1-f)(4n)^{-1}\left\{f_Y(Me_Y)\right\}^{-2}. \tag{2.1.6}$$

It is shown in Appendix A that $\hat{M}e_{YR} - Me_Y$ is asymptotically normal with mean zero and variance

$$n^{-1}(1-f)[\frac{1}{4}\left\{f_Y(Me_Y)\right\}^{-2} + \frac{1}{4}(Me_Y/Me_X)^2\left\{f_X(Me_X)\right\}^{-2} - \tag{2.1.7}$$

$$+2(Meee_Y/Me_X)\left\{f_Y(Me_Y)f_X(Me_{X)}\right\}^{-1}(P_{11}-\frac{1}{4})$$

Consequently, $\hat{M}e_{YR}$ is asymptotically more efficient than $\hat{M}e_Y$ if

$$\rho_c > \frac{1}{2}\left[\left\{f_X(Me_X)\right\}^{-1}Me_X^{-1}\right] / \left[\left\{f_Y(Me_Y)\right\}^{-1}Me_Y^{-1}\right],$$

where $\rho_c = 4\left(P_{11} - \frac{1}{4}\right)$ goes from -1 to 1 as P_{11} increases from 0 to 1/2. This condition is analogous to the condition under which the ratio estimator is superior to the sample mean. Since P_{11} is the proportion of units in the population with $X \leq Me_X$ and $Y \leq Me_Y$, it can be regarded as a measure of concordance.

It is also shown in Appendix A that $\hat{M}e_{YP}$ and $\hat{M}e_{YS}$ have the same asymptotic distribution which is Normal with mean Me_Y and variance

$$2\left\{f_Y(Me_Y)\right\}^{-2}(1-f)P_{11}(1-2P_{11})n^{-1}. \tag{2.1.8}$$

If, for variables X and Y, the condition $X \le Me_X$ if and only if $Y \le Me_Y$ holds, then $P_{11} = \frac{1}{2}$ in which case this variance expression is reduced to zero and both $\hat{M}e_Y$ - Me_Y and $\hat{M}e_{YS} - Me_Y$ are of order $o_p(n^{-1/2})$. When the ordering of the Y_S and that of the X_S in the population agree closely with each other, P_{11} is close to $\frac{1}{2}$ and the asymptotic variance is small. Thus the efficiencies of the estimators $\hat{M}e_{YP}$ and $\hat{M}e_{YS}$ do not rely on any linearity assumption between X and Y. Comparing the expression 2.1.8 with expression 2.1.6, we conclude that both $\hat{M}e_{YP}$ and $\hat{M}e_{YS}$ are asymptotically more efficient than the sample median, since

$$\frac{1}{4} - 2P_{11}(1-2P_{11}) = \frac{1}{4}(4P_{11}-1)^2 \ge 0.$$

The common asymptotic variance of $\hat{M}e_{YP}$ and $\hat{M}e_{YS}$ involves only $f_Y(Me_Y)$ and P_{11}, the latter being consistently estimated by p_{11}. The value $f_Y(Me_Y)$ can be estimated by applying standard methods such as the kernel or the nearest neighbor method of density estimation specifically at the sample median $\hat{M}e_Y$. The resulting estimator of $f_Y(Me_Y)$ together with p_{11} can be used to substitute for $f_Y(Me_Y)$ and P_{11} to yield a consistent estimator of the asymptotic variance of $\hat{M}e_{YP}$ and $\hat{M}e_{YS}$.

2.2. Median estimation without auxiliary information

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 [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 $x > 0$.

Williams [1946] 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 .

Pólya [1945] 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

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

Applying (1) to the upper and lower bounds, Chu [1955] obtained

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

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

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

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

$$\phi(t) = \int\limits_0^t (1/\sqrt{2\pi})exp(-\frac{1}{2}x^2)dx.$$

For more details see the work of Chu [1955].

The application will be presented in Chapter 3.

2.3. Median estimation in the presence of auxiliary information

The estimators proposed by Kuk and Mak [1989], are applicable in situations where only the population median or a grouped frequency distribution of the auxiliary variable is known.

Suppose the population studied consists of N units. Each of these N units are the values of the survey variable Y of interest and an auxiliary variable X. It is first assumed that only the median Me_X of X is known and that Me_Y is to be estimated on the basis of a simple random sample S_n of size n.

Let $(X_1, Y_1), ..., (X_n, ..., Y_n)$ be the associated values of the variables X and Y for the units in S_n. In the absence of the X_i, the sample median $\hat{Me}_Y$ is a natural estimator of M_Y. When the values of the auxiliary variable X_i are available, a natural modification of the ratio estimator for estimating population mean is

$$\hat{Me}_{YP} = \hat{Me}_Y Me_X / \hat{Me}_X.$$

Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$be the ordered Y values in S_n. Also, let i_0 be the integer such that $Y_{(i_0)} \leq M_Y \leq Y_{(i_0+1)}$ and $p = i_0/n$ be the proportion of the Y_S in the sample that are less than or equal to Me_Y. Thus, Me_Y is approximately the sample pth quantile $\hat{Q}_Y(p)$. The sample median $\hat{Me}_Y$ can be viewed as the special estimator $\hat{Q_Y}(\hat{p})$ with $\hat{p} = \frac{1}{2}$.

The estimator defined attempts to utilize the X_i to construct a $\hat{p}$ with smaller expected squared error $E(\hat{p}-p)^2$ and consequently smaller variance for estimating Me_Y. Consider now the cross-classification

	$X \leq Me_X$	$X > Me_X$
$Y \leq Me_Y$	P_{11}	P_{12}
$Y > Me_Y$	P_{21}	P_{22}

where for instance P_{11} denotes the proportion of units in the population with $X \leq Me_X$ and $Y \leq Me_{Y.}$ Let n_X be the number of units in S_n with $X \leq Me_X$. Then, if the P_{ij} are known, we can estimate p by

$$\begin{aligned} \hat{p} &= n^{-1}\left\{n_X P_{11}/P_{.1} + (n - n_X)P_{12}/P_{.2}\right\} \\ &\simeq (2/n)\left\{n_X P_{11} + (n - n_x)(\frac{1}{2} - P_{11})\right\}, \end{aligned}$$

where $P_{.j} = P_{1j} + P_{2j} \simeq \frac{1}{2}$, for $j = 1$ and $j = 2$. In practice, the P_{ij} are usually unknown but can be estimated by $\hat{p_{ij}}$ based on a similar cross-classification of the sample. Thus $\hat{p}_{11}$ represent the proportion of units in the sample with $X \leq \hat{Me}_X$ and $Y \leq \hat{Me}_Y$. Substituting the $\hat{p_{ij}}$ for the P_{ij}in the expression for $\hat{p}_0$, we have the estimator of p:

$$\begin{aligned} & n^{-1}\left\{n_X p_{11}/p_{.1} + (n - n_X)p_{12}/p_{.2}\right\} \\ \simeq\ & (2/n)\left\{n_X p_{11} + (n - n_x)(\frac{1}{2} - p_{11})\right\}. \end{aligned}$$

An estimator of Me_Y is alternatively given by

$$\hat{M}e_{YP} = \hat{Q}_Y(\hat{p}_1)$$

and will be referred to as the 'position estimator' since it is essentially based on an estimate of the position of Me_Y in the ordered sample values.

The cross-classification used in constructing $\hat{M}e_{YP}$ motivates yet another way of estimating Me_Y.

The idea is essentially to invert an improved estimate of F_Y based on post-stratification; for any value y, let $F_{Y1}(y)$ be the proportion among those units in the sample with $X \leq Me_X$ that have Y values less than equal to y. Similarly $\tilde{F}_{Y2}(y)$ is the proportion among those with $X > Me_X$. Then $F_Y(y)$ can be estimated by

$$\tilde{F}_Y(y) = N^{-1}N_X\tilde{F}_{Y1}(y) + N^{-1}(N - N_X)\tilde{F}_{Y2}(y)$$

$$\simeq \frac{1}{2}\left\{\tilde{F}_{Y1}(y) + \tilde{F}_{Y2}(y)\right\},$$

where N_X is the number of units in the population with $X \leq Me_X$. Note that $\tilde{F}_Y(y)$ is a distribution function which, for each Y_i, put probability mass $(2n_X)^{-1}$ at Y_i if $X_i \leq Me_X$ and $\{2(n - n_X)\}^{-1}$ if $X_i > Me_X$. Thus an estimator of Me_Y is given by

$$\hat{M}e_{YS} = inf\left\{y : \tilde{F}_Y(y) \geq \frac{1}{2}\right\}$$

and will be referred to as the ' stratification estimator'.

2.3.1. Ratio estimators.

Estimating a ratio.

We consider a finite population of size N in which two characters X and Y assume positive values. Suppose we know the true average. Consider a sample, with or without replacement, of n elements that gives rise to pairs of values:

$$(x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n).$$

We will be interested in the quantity

$$\bar{y}_R = Y_T/X_T = \bar{Y}/\bar{X} \tag{2.3.1}$$

which we will refer to as the population ratio.

There are various possible approaches to estimate $\bar{y}_R$. Two immediately obvious ones are to use the sample average ratio or the ratio of the sample averages.

Specifically, these are

$$\bar{y}_{R_1} = \frac{1}{n}\sum_{i=1}^{n}(y_i/x_i)$$

and

$$\bar{y}_{R_2} = \bar{y}/\bar{x} = y_T/x_T,$$

respectively. We will examine some of the sampling properties of $\bar{y}_{R_1}$and $\bar{y}_{R_2}$.

(a) Consider the population of values $R = Y/X$. This has population mean $\bar{R}$ and variance S_R^2. Since $\bar{y}_{R_1}$ is a sample mean from a sr sample, it has expected value $\bar{R}$ and variance $(1-\frac{n}{N})S_R^2/n$.

But typically $\bar{R}$ is not the same as R, so we have

$$\begin{aligned} bias(\bar{y}_{R_1}) &= \bar{R} - R \\ &= \tfrac{1}{N}\textstyle\sum_i R_i \quad -Y_T/X_T \\ &= \tfrac{1}{N}\textstyle\sum_i (Y_i/X_i) \quad -\sum_i^N Y_i/\ \sum_i^N X_i \\ &= -\tfrac{1}{X_T}\textstyle\sum_1^N R_i(X_i - \bar{X}). \end{aligned} \tag{2.3.2}$$

This features the covariance $Cov(R,X) = \sum_1^N R_i(X_i-\bar{X})/(N-1)$ between R and X.

Thus, noting that the mean square error is the sum of the variance and the square of the bias, we have

$$MSE(\bar{y}_{R_1}) = \left(1-\frac{n}{N}\right)S_R^2/n + (N-1)^2 Cov(R,X)/(X_T)^2.$$

We have the usual unbiased variance estimator $\sum_1^n (r_i-\bar{r})^2/(n-1)$ available for S_R^2 and we can obtain an unbiased estimator of the covariance $Cov(R,X)$ in the form:

$$\sum_1^n r_i(x_i-\bar{x})/(n-1) = n(\bar{y}-\bar{r}\bar{x})/(n-1).$$

We can estimate the bias and the MSE of $\bar{y}_{R_1}$ by means of

$$-(N-1)n(\bar{y}-\bar{y}_{R_1}\bar{x})/[(n-1)X_T]$$

and

$$\left(1-\frac{n}{N}\right)\sum_1^n (r_i-\bar{y}_{R_1})^2/n + (N-1)^2 n^2(\bar{y}-\bar{y}_{R_1}\bar{x})^2[(n-1)^2X_T^2]$$

respectively, provided X_T is know. If X_T were known, we could correct $\bar{y}_{R_1}$ for estimated bias obtaining a modified estimator

$$\bar{y}^{'}_{R_1} = \bar{y}_{R_1} + (N-1)n(\bar{y} - \bar{y}_{R_1}\bar{x})/[(n-1)X_T] \tag{2.3.3}$$

this is the Hartley-Ross estimator.

(b) Let us consider $\bar{y}_{R_2}$

This estimator is more widely used. Although still biased in small samples, the bias and mean square error tend to be lower than the bias of $\bar{y}_{R_1}$. In large samples the bias becomes negligible and the distribution of $\bar{y}_{R_1}$ tends to normality, thus making it possible to draw inferences based on a Normal distribution with appropriate variance.

Let us start again with the bias. We have

$$\bar{y}_{R_2} - R = (\bar{y} - R\bar{x})/\bar{x}$$

and taking a Taylor series expansion about the population mean $\bar{X}$ gives

$$\bar{y}_{R_2} - R = \frac{\bar{y} - R\bar{x}}{\bar{X}}\left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)^{-1}.$$

As an approximation to the bias we can take the firtst two terms to obtain

$$E(\bar{y}_{R_2}) - R = E\left(\frac{\bar{y} - R\bar{x}}{\bar{X}}\right) - \frac{1}{X^2}E\left[(\bar{y} - R\bar{x})(\bar{x} - \bar{X}\right].$$

The leading term is zero since $E(\bar{y} - R\bar{x}) = \bar{Y} - R\overline{X} = 0$. Furthermore,

$$E\left[\bar{y}\left(\bar{x} - \bar{X}\right)\right] = Cov(\bar{y}, \bar{x}) = \left(1 - \frac{n}{N}\right)Cov(X, Y) = \left(1 - \frac{n}{N}\right)\rho_{XY}S_Y S_X/n,$$

where ρ_{YX} is the correlation between Y and X. Thus we find, as an approximation to the bias of $\bar{y}_{R_2}$,

$$E(\bar{y}_{R_2}) - R = \frac{(1 - n/N)}{n\bar{X}^2}\left(RS_X^2 - \rho_{YX}S_Y S_X\right)$$

which can be small if ρ_{YX} is close in value to RS_X/S_Y.

This is equivalent to saying that the regression of Yon X is linear and through the origin, or that Y and X are roughly proportional to each other.

Suppose we consider large samples, utilizing asymptotic results, we find the following approximate results:

$$E(\bar{y}_{R_2}) = \bar{Y}/\bar{X} = Y_T/X_T$$

and

$$Var(\bar{y}_{R_2}) = \frac{1-f}{n\bar{X}^2}\sum_{i=1}^{N}\frac{(Y_i - RX_i)^2}{N-1}$$

where f is the sampling fraction n/N.

Ratio estimator of a population total or mean.

Suppose we want to known the estimate of total population.

Suppose Y_i denotes expenditure on recreational facilities for authority i, X_i denotes the number of inhabitants in the authority, and we sample both mesaures simultaneously and at random from the whole population, to obtain a sr sample of size n: $(y_1, x_1), ..., (y_n, x_n)$. The total number of inhabitants for the whole population X_T, is likely to be known fairly accurately. However we could estimated X_T from the sample by means of the estimator

$$x_T = N\bar{x},$$

where $\bar{x}$ is the sr sample mean. Similarly we could estimate the total expenditure Y_T by

$$y_T = N\bar{y}.$$

The estimate x_T has not interest in its own right, but it has the important advantage that by comparing it with the population characteristic X_T we can informally assess the representativeness of the sample. If x_T is very much less than X_T, then in view of the rough proportionality of Y_i and X_i we could conclude that y_T is likely to understimate Y_T; if x_T is too large, y_T is also likely to be too large. If the proportionality relationship were exact we would have

$$Y_i = RX_i \tag{2.3.4}$$

where R is the population ratio, Y_T/X_T or $\bar{Y}/\bar{X}$.

Thus,

$$Y_T = RX_T$$

and we could estimate Y_T by replacing R with the sample estimate, $\bar{y}_{R_2}$, to obtain an estimate of the population total, Y_T,in the form

$$y_{TR} = \bar{y}_{R_2}X_T = \frac{X_T}{x_T}y_T. \tag{2.3.5}$$

This estimator is called the sample ratio estimator of the population total.

If interest centeres on the population mean $\bar{Y}$, rather than the total Y_T,then similar arguments support the use of the ratio estimator of the population mean

$$\hat{y}_R = \frac{\bar{X}}{\bar{x}}\bar{y}. \tag{2.3.6}$$

Considering 2.3.6 , the approximate variance is

$$Var(\hat{y}_R) = \frac{1-f}{n}\sum_{i=1}^{N}\frac{(Y_i - RX_i)^2}{N-1} = \frac{1-f}{n}\left(S_Y^2 - 2RS_{XY} + R^2S_X^2\right).$$

That is

$$Var(\hat{y}_R) = \frac{1-f}{n}\left(S_Y^2 - 2R\rho_{YX}S_YS_X + R^2S_X^2\right) \tag{2.3.7}$$

where $\rho_{YX} = S_{YX}/S_YS_X$ is the population correlation coefficient. If the exact relationship 2.3.4 held, then $Var(\hat{y}_R)$ would be zero; in practice this will not be so, but $Var(\hat{y}_R)$ is clearly going to become smaller, the larger the positive correlation between Y and X in the population.

For estimating Y_T we have an analogous result for y_{TR}. It is asymptotically unbiased, and has large sample variance

$$\frac{N^2(1-f)}{n}\sum_{i=1}^{N}\frac{(Y_i - RX_i)^2}{N-1}$$

or

$$\frac{N^2(1-f)}{n}\left(S_Y^2 - 2R\rho_{YX}S_YS_X + R^2S_X^2\right).$$

Consider the following model

$$Y_i = RX_i + e_i, \tag{2.3.8}$$

with $\sum_x E_i = 0$, where $\sum_x$ denotes summation over all subscripts i for which $X_i = x$.

In this case $\bar{Y} = R\bar{X}$ and in a sample of size n

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = R + \frac{\bar{e}}{\bar{x}}$$

where $\bar{e}$ is the sample mean of the E values in the sample.

The conditional expectation

$$E(\bar{e} \mid x_1, ..., x_n) = 0$$

for this model, so $E(\hat{R}) = R$and we conclude that $\hat{R}$ is unbiased for all sample sizes [Barnett, 1991].

2.3.2. Median estimation using the ratio's estimator .

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

$$\hat{Y}_R = \frac{\hat{M}e(Y) * Me(X)}{\hat{M}e(X)} \tag{2.3.9}$$

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 estimators. 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

$$\frac{1}{\hat{M}e_X} = \frac{1}{Me_X + (\hat{M}e_X - Me_X)} = \frac{1}{Me_X}\frac{1}{1 + \frac{\hat{M}e_X - Me_X}{Me_X}}$$

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

$$\simeq \frac{1}{Me_X}\left(1 - \frac{\hat{M}e_X - Me_X}{Me_X}\right)$$

then

$$\left(\frac{\hat{M}e_Y}{\hat{M}e_X} - \frac{Me_Y}{Me_X}\right) \cong \left(\hat{M}e_Y - \frac{Me_Y}{Me_X}\hat{M}e_X\right) * \frac{1}{Me_X}\left(1 - \frac{\hat{M}e_X - Me_X}{Me_x}\right) =$$

$$= \frac{1}{Me_X}\left(Me_Y - \frac{\hat{M}e_Y}{Me_X}\;\hat{M}e_X\right) - \frac{1}{Me_X}\left(\hat{M}e_Y - \frac{Me_Y}{Me_X}\hat{M}e_X\right)\left(\frac{\hat{M}e_X - Me_X}{Me_X}\right).$$

Passing to the squares

$$\left(\frac{\hat{M}e_Y}{\hat{M}e_X} - \frac{Me_Y}{Me_X}\right)^2 =$$

$$= \frac{1}{Me_X^2}\left[\left(\hat{M}e_Y - Me_Y\right) - \left(\frac{Me_Y}{Me_X}\hat{M}e_X - Me_Y\right)\right]^2 +$$

$$+\frac{2}{Me_X^2}\left(\hat{M}e_Y - \frac{Me_Y}{Me_X}\hat{M}e_X\right)^2\left(\frac{\hat{M}e_X - Me_X}{Me_X}\right)^2$$

we define the expected value

$$E\left(\frac{\hat{M}e_Y}{\hat{M}e_X}-\frac{Me_Y}{Me_X}\right)^2 \simeq \frac{1}{Me_X^2}[E(\hat{M}e_Y-Me_y)^2+$$

$$+\frac{Me_Y^2}{Me_X^2}E(\hat{M}e_X-Me_X)^2-2\frac{Me_Y}{Me_X}E(\hat{M}e_Y-Me_Y)(\hat{M}e_X-Me_X)]+$$

$$+\frac{1}{Me_X^2}E\left[(\hat{M}e_Y-Me_Y)^2(\hat{M}e_X-Me_X)^2\right].$$

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

$$MSE(\hat{M}e_Y^R)=MSE\left(Me_X\frac{\hat{M}e_Y}{\hat{M}e_X}\right)=Me_X^2 * MSE\left(\frac{\hat{M}e_Y}{\hat{M}e_X}\right)=$$

$$=E\left[\left(\hat{M}e_Y-Me_Y\right)^2\right]+\frac{Me_Y^2}{Me_X^2}E(\hat{M}e_X-Me_X)^2-2\frac{Me_Y}{Me_X}[Cov(\hat{M}e_Y,\hat{M}e_X)+$$

$$+(Me_Y-E(\hat{M}e_Y))(Me_X-E(\hat{M}e_X))]$$

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.

2.3.3. Median Regression.

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

$$\hat{y}=a+bx \tag{2.3.10}$$

that, given N points $(x_{i,}y_i)$, make the minimum sum of the absolute values S of the residuals r_i:

$$S=\sum_{i=1}^N \mid y_i-\hat{y}_i \mid = \sum_{i=1}^N \mid y_i-a-bx_i \mid. \tag{2.3.11}$$

The problem has been resolved geometrically by Boscovich and subsequently formalized by Laplace (1786) , in the case it requiring 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), 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.

- **The restricted problem**

Given a set of points $\{x_i, y_i\} i = 1, 2, ..., n$ find the equation of the line 2.3.10 , through any point $(x^*$,$y^*)$, not necessarily one of the given set, such that

$$S = \textstyle\sum_{i=1}^{N} \mid y_i - a - bx_i \mid$$

$$= \textstyle\sum_{i=1}^{N} \mid y_i - bx_i - (y^* - bx^*) \mid \tag{2.3.12}$$

$$\textstyle\sum_{i=1}^{N} \mid (y_i - y^*) - b(x_i - x^*) \mid .$$

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

$$x_i^{'} = x_i - x^* \tag{2.3.13}$$

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

for $i = 1, 2, ..., N$.

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

$$y^{'} = bx \tag{2.3.14}$$

such that

$$S = \sum \mid y_i - y_i^{'} \mid \tag{2.3.15}$$

is minimum, where

$$\hat{y}^{'} \equiv bx_i^{'}. \tag{2.3.16}$$

S may be written

$$S = \sum_{i=1}^{n} \mid y_i^{'} - \hat{y}_i^{'} \mid = \sum_{i=1}^{n} \mid y_i^{'} - bx_i^{'} \mid . \tag{2.3.17}$$

Hence, the problem becomes finding b in 2.3.17 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 2.3.13.

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

$$\hat{y} = \left(\frac{y_i^{'}}{x_i^{'}}\right) x \tag{2.3.18}$$

in the transformed coordinates, or

$$y^{'} - y^* = \left(\frac{y_i^{'}}{x_i^{'}}\right)(x - x^*) \tag{2.3.19}$$

in the original coordinates.

To illustrate the method proposed by Karst [1958] we shall find the line of best fit for data shown in columns (1) and (2) of Table 6. We shall restrict this line to pass throught the centroid of the data. The various columns of Table 6 summarize the essential calculation necessary to find the slope of the desired line. The detailed steps of the analysis are given below the table.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
x_i	y_i	$x_i^{'} = x_i - \bar{x}$	$y_i^{'} = y_i - \bar{y}$	$\frac{y_i^{'}}{x_i^{'}}$	Rank	$\tilde{x}_j$
65	68	-1.67	0.417	-0.25	2	0.33
63	66	-3.67	-1.583	0.431	6	-1.67
67	68	0.33	0.417	1.264	10	3.33
64	65	-2.67	-2.853	0.967	8	2.33
68	69	1.33	1.417	1.065	9	-4.67
62	66	-4.67	-1.583	0.339	5	-3.67
70	68	3.33	0.417	0.125	3	4.33
66	65	-0.67	-2.583	3.855	12	-2.67
68	71	1.33	3.417	2.57	11	1.33
67	67	0.33	-0.583	-1.767	1	0.33
69	68	2.33	0.417	0.179	4	1.33
71	70	4.33	2.417	0.558	7	0.67
800	811					

TABLE 6. Determinig the slope of the line

STEPS OF ANALYSIS OF TABLE *66.* First we calculate the arithmetic means of X and Y:

$$\bar{x} = \frac{800}{12} = 66.67$$

$$\bar{y} = \frac{811}{12} = 67.583.$$

1. Columns (1), (2) are the recordings of raw data.
2. Columns (3), (4) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using
$$x_i^{'} = x_i - \bar{x}$$
$$y_i^{'} = y_i - \bar{y}.$$
3. Column (5) is the ratio $\frac{y_i^{'}}{x_i^{'}}$; these numbers are the b values at the minimum points of the individual curves of the terms $\mid y_i^{'} - bx_i^{'} \mid$.
4. Column (6) ranks the data of column (5) in ascending algebraic order.
5. By adding the absolute values of column (3), we obtain $\sum \mid x_i^{'} \mid = 26.66$. This is the magnitude of the extreme left and right slopes of the function $S = \sum_{i=1}^{n} \mid y_i^{'} - bx_i^{'} \mid$.
6. The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until a change in sign is obtained. Thus:

$$\begin{array}{rr}
-\sum \mid x_i' \mid: & \text{-26.66} \\
2 \mid \tilde{x}_1 \mid: & 0.66 \\ \cline{2-2}
 & \text{-26} \\
2 \mid \tilde{x}_2 \mid: & 3.34 \\ \cline{2-2}
 & \text{-22.66} \\
2 \mid \tilde{x}_3 \mid: & 6.66 \\ \cline{2-2}
 & \text{-16} \\
2 \mid \tilde{x}_4 \mid: & 4.66 \\ \cline{2-2}
 & \text{-11.34} \\
2 \mid \tilde{x}_5 \mid: & 9.34 \\ \cline{2-2}
 & \text{-2} \\
2 \mid \tilde{x}_6 \mid: & 7.34 \\ \cline{2-2}
 & 5.34
\end{array}$$

Since the index i=6 effects the change in sign, $b = y_2^{'}/x_2^{'}$=0.431 is the slope of the line of best fit,

$$\hat{y} = 0.431x.$$

The line of the best fit in terms of the original variables is

$$y - 67.583 = 0.431(x - 66.67)$$

or

$$\hat{y} = 0.431x + 38.848.$$

- **The unrestricted problem.**

Suppose now the case where it is not required that the straight line pass through a predetermined point (x_i^*, y_i^*), in addition to the angular coefficient, it is also necessary to determine the intercept , that is, we want to determine the values of parameters that minimize the sum :

$$S = \sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid = \sum_{i=1}^{n} \mid y_i - bx_i - a \mid . \tag{2.3.20}$$

Karst (1958) proposes the iterative procedure which will be shown to lead to the values of b and a associated with the absolute minimum of S.

(1) Choose any point (x_1, y_1) of the given data and determine its associated local minimum line

$$\hat{y} = b_1 x + a_1. \tag{2.3.21}$$

(2) This line passes through at least one other point (x_2, y_2) of the given data. Using (x_2, y_2)as a reference point, find its local minimum line

$$\hat{y} = b_2 x + a_2. \tag{2.3.22}$$

It will be shown that the sum of the absolute deviations S_2 associated with 2.3.22 is less than S_1 associated with 2.3.21.

(3) Line 2.3.22 passes through another point (x_3, y_3) which, in turn, is used as a reference point to determine

$$\hat{y} = b_3 x + a_3. \tag{2.3.23}$$

with $S_3 < S_2 < S_1$.

(4) This procedure is repeated until a point (x_t, y_t) is reached such that the local minimum line

$$\hat{y} = b_t x + a_t. \tag{2.3.24}$$

does not pass through a new point (x_{t+1}, y_{t+1}) but rather reflects back to the previous point (x_{t-1}, y_{t-1}). This line 2.3.24 is then the absolute minimum line of the data.

We illustrate the procedure with an example, taken from the work of RADAELLI [2004]

The data are presented in Zenga [1988] (pag.320) and refer to the numeber of subscribers (Y) to a journal for the years 1979 to 1985.

Years	x_i	y_i
1979	0	6284
1980	1	7603
1981	2	8591
1982	3	9278
1983	4	10206
1984	5	10614
1985	6	11569

Determine the equation of the line :

$$\hat{y} = bx + a \tag{2.3.25}$$

that minimizes :

$$S = \sum_{i=1}^{6} \mid y_i - \hat{y}_i \mid = \sum_{i=1}^{6} \mid y_i - bx_i - a \mid . \tag{2.3.26}$$

We choose as starting point the values of 2003 ($x_1 = 2; y_1 = 8591$). Applying the procedure described in paragraph 2.3.3 to the remaining point, we obtain the equation of the straight line passing through the point (2 ; 8591) that minimizes the sum of absolute value :

$$\hat{y} = 807.5x + 6979 \tag{2.3.27}$$

in correspondence of which we have $S = S_1 = 1644,5$. The line 2.3.27 passes through the point of coordinates $(4; 10206)$that becomes the new reference point for determination of the new angular coefficient. Again applying the procedure of paragraph 2.3.3 we obtain the line passing through the point (4;10206) that minimizes S:

$$\hat{y} = 867,\bar{6} + 6735,\bar{3}. \tag{2.3.28}$$

The value of S for 2.3.28 is :

$$S_2 = 1464 < 1644,5 = S_1.$$

The line 2.3.28 passes through the point of coordinates (1;7603) that does not coincide with the previous point $(2; 8591)$, therefore, we must find the new line through point $(1; 7603)$,which minimizes S. This line is :

$$\hat{y} = 793,2x + 6809,8 \tag{2.3.29}$$

in correspondence of which we have $S = S_3 = 1194,4 < 1464 = S_2$. The line 2.3.29 passes through the point $(6; 11569)$ that is different from the previous point $(4; 10206)$ and therefore the algorithm continues.

Then we obtain for the line through ($x_4 = 6; y_4 = 11569$) that minimizes S :

$$\hat{y} = 793,2x + 6809,8. \tag{2.3.30}$$

The line 2.3.30 coincides with 2.3.29, infact 2.3.30 passes through the point of the previous reference $(1; 7603)$ and the algorithm terminates.

2.3.30 is the equation of the line that minimizes the sum of the absolute value of the deviations of the $n = 7$ points. The minimum value of this sum is :

$$S = S_4 = S_3 = 1194,4.$$

For more details see the work of Karst [1958] and Radaelli [2002].

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

$$E\left(\hat{M}e_{MR}-Me_Y\right)^2 = E\left(\hat{M}e_Y + b(Me_X - \hat{M}e_X) - Me_Y\right)^2 \tag{2.3.31}$$

$$= E[(\hat{M}e_Y - Me_Y)^2 + b^2(\hat{M}e_X - Me_X)^2 - 2b(\hat{M}e_Y - Me_Y)(\hat{M}e_X - Me_X)]$$
$$= E(\hat{M}e_Y - Me_Y)^2 + b^2 E(\hat{M}e_X - Me_X)^2 - 2bCov(\hat{M}e_Y, \hat{M}e_X).$$

For application, refer to Chapter 3.

2.3.4. Linear regression . Suppose that the principal variable Y may be expressed as a linear function of the auxiliary variable X, that is

$$Y_i = a + bx_i + e_i$$

where e_i is a random variable due to errors of detection and to the dependence of Y_i by other variables not specified.

It is assumed that the Y_i are independent from each other and that their expected value and variance are :

$$E(Y_i) = a + bx_i$$

and

$$Var(Y_i) = \sigma_y^2.$$

Suppose we know the true mean μ_x of the random variable X. We extract a sample of size n. Based on sample data, we obtain the pairs (x_i, y_i), $i = 1, ..., n$ observing on each unit extracted the character X and the character Y.

Watson, in 1937, used the regression of the weight of leaves on the leaves of certain plants, to determine the weight and the area given every leaf belongs to a small sample. The average area of leaves was then adjusted by regressing the weight of the leaves.

The regression method is justified by the fact that it is very easy to detect the weight of the leaves while determination of the area requires more time. This example suggests that the regression method is useful if it is easy to detect information about μ_x, while it is expensive to detect the characteristic Y on the units.

To proceed with estimate of $\mu_y$by the regression method , we distinguish two cases based, or not, on the knowledge of the angular coefficient [Pollastri, 1997].

- Estimate of the mean by the regression method .

Assume that the connection between X and Y is

$$E(Y_i) = a + b_0 x_i$$

where b_0 is the value chosen for b.

The estimate with the regression method is given by :

$$\bar{y}_{lr} = \bar{y} + b_0(\mu_x - \bar{x}).$$

The estimator $\bar{Y}_{lr}$ is unbiased

$$E(\bar{Y}_{lr}) = E(\bar{Y}) + b_0 E(\mu_x - \bar{X}) = \mu_y$$

and has variance

$$Var(\bar{Y}_{lr}) = \frac{\sigma_y^2}{n}(1 - \rho^2).$$

According to the pairs of values obtained from the sample, b and a are estimated by the method of least squares

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.3.32}$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

2.3.4.1. 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 2.3.32 and the population median of the auxiliary variable X , the estimate of the median of Y using the regression method is :

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

2.3.5. Quantile Regression.

Quantiles . Any real-valued random variable X may be characterized by its (right-continuous) distribution function

$$F(x) = P(X \leq x), \tag{2.3.33}$$

whereas for any $0 < \tau < 1$,

$$F^{-1}(\tau) = inf\,\{x : F(x) \geq \tau\} \tag{2.3.34}$$

is called the τth quantile of X.

The quantiles arise from a simple optimization problem. Consider a simple decision theoretic problem: a point estimate is required for a random variable with (posterior) distribution function $F(\cdot)$. If loss is described by the piecewise linear function

$$\rho_\tau(u) = u(\tau - I(u < 0)) \tag{2.3.35}$$

for some $\tau\in (0, 1)$, find $\hat{x}$ to minimize expected loss. We seek to minimize

$$E\rho_\tau(X - \hat{x}) = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \int_{\hat{x}}^{\infty} (x - \hat{x})dF(x). \tag{2.3.36}$$

Differentiating with respect to $\hat{x}$, we have

$$0 = (1 - \tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau. \tag{2.3.37}$$

Since $F(\cdot)$is monotone, any element of $\{x : F(x) = \tau\}$ minimizes expected loss. When the solution is unique, $\hat{x} = F^{-1}(\tau)$; otherwise, we have an " interval of τth quantiles" from which the smallest element must be chosen - to adhere to the convention that the empirical quantile function be left-continuous. It is natural than an optimal point estimator for asymmetric linear loss should lead us to the quantiles. In the asymmetric case of absolute value loss it is well known to yield the median. When loss is linear and asymmetric, we prefer a point estimate more likely to leave us on the flatter of the two branches of marginal loss. Thus, for example, if an underestimate is marginally three times more costly than an overestimate, we will choose $\hat{x}$ so that $P(X \leq \hat{x})$is three times greater than $P(X > \hat{x})$ to compensate. That is, we will choose $\hat{x}$ to be the 75th percentile of $F(\cdot)$.

When $F(\cdot)$ is replaced by the empirical distribution function

$$F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x), \tag{2.3.38}$$

we may still choose $\hat{x}$ to minimize expected loss:

$$\int \rho_\tau(x-\hat{x})dF_n(x) = n^{-1}\sum_{i=1}^{n}\rho_\tau(x_i-\hat{x}) \tag{2.3.39}$$

and doing so now yields the τth sample quantile.

The problem of finding the τth sample quantile, which can be written as

$$\min_{\xi\in\mathbf{R}}\sum_{i=1}^{n}\rho_\tau(y_i-\xi), \tag{2.3.40}$$

may be reformulated as a linear program by introducing $2n$ artificial, or "slack", variables $\{u_i, v_i : 1, \ldots, n\}$ to represent the positive and negative parts of the vector of residuals. This yields the new problem

$$\min_{(\xi,u,v)\in\mathbb{R}\times\mathbb{R}_+^{2n}}\left\{\tau 1_n^\top u + (1-\tau)1_n^\top v \mid 1_n\xi + u - v = y\right\}, \tag{2.3.41}$$

where 1_n denotes an n-vector of 1. Clearly, in (2.3.41) we are minimizing a linear function of a polyhedral constraint of the intersection of the $(2n+1)-$dimensional hyperplane determined by linear equality constraints and the set $\mathbb{R}\times\mathbb{R}_+^{2n}$.

2.3.5.1. Introduction to quantile regression.

That the quantiles may be expressed as the solution to a simple optimization problem leads, naturally, to more general methods of estimating models of conditional quantile functions. Least squares offers a template for this development.

Knowing that the sample mean solves the problem

suggests that, if we are willing to express the conditional mean of y given x as $\mu(x) = x^\top\beta$, then β may be estimated by solving

$$\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}(y_i - x_i^\top\beta)^2. \tag{2.3.42}$$

Similarly, since the τth sample quantile, $\hat{\alpha}(\tau)$, solves

$$\min_{\alpha\in\mathbb{R}}\sum_{i=1}^{n}\rho_\tau(y_i-\alpha), \tag{2.3.43}$$

we are led to specifying the τth conditional quantile function as $Q(\tau \mid x) = x^\top\beta(\tau)$, and to consideration of $\hat{\beta}(\tau)$ solving

$$\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}\rho_\tau(y_i - x_i^\top\beta). \tag{2.3.44}$$

This is the germ of the idea elaborated by Koenker and Bassett (1978) [Koenker and Bassett Jr, 1978].

Quantile regression problem 2.3.44 may be reformulated as a linear program as in 2.3.41 :

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}^{2n}_{+}} \left\{\tau 1_n^\top u + (1-\tau)1_n^\top v \mid X\beta + u - v = y\right\}, \tag{2.3.45}$$

where X now denotes the usual n by p regression design matrix. We have split the residual vector $y - X\beta$ into its positive and negative parts, and so we are minimizing a linear function on a polyhedral constraint set, and most of the important properties of the solutions, $\hat{\beta}(\tau)$, which we call "regression quantiles".

Koenker and Bassett now deal with regression quantile introducing the central special case, which is the median regression estimator that minimizes a sum of absolute errors.

We can say that a student scores at the τth quantile of a standardized exam if he/she performs better than the proportion τ, of the reference group of students, and worse than the proportion $(1-\tau)$. Thus, half of the students perform better than the median student, and half perform worse. Similarly, the quartiles divide the population into four segments with equal proportions of the reference population in each segment. The quantiles, or percentiles refer to the general case. Quantile regression seeks to extend these ideas to the estimation of conditional quantile functions, models in which quantiles of the conditional distribution of the response variable are expressed as functions of observed covariates. To accomplish this task Koenker and Hallock [2001] needed a new way to define the quantiles.

Quantiles seem inseparably linked to the operations of the ordering and sorting that are generally used to define them. Koenker and Hollack define quantiles through a simple alternative as an optimization problem. Just as we can define the sample mean as the solution to the problem of minimizing a sum of squared residual, we can define the median as the solution to the problem of minimizing a sum of absolute residuals. Regarding the other quantiles, if the symmetric absolute value function yields the median, perhaps we can simply tilt the absolute value to produce the other quantiles. This logic suggests solving

$$\min_{\xi\in\Re} \sum \rho_\tau(y_i - \xi) \tag{2.3.46}$$

where the function $\rho_\tau(\cdot)$ is illustrated

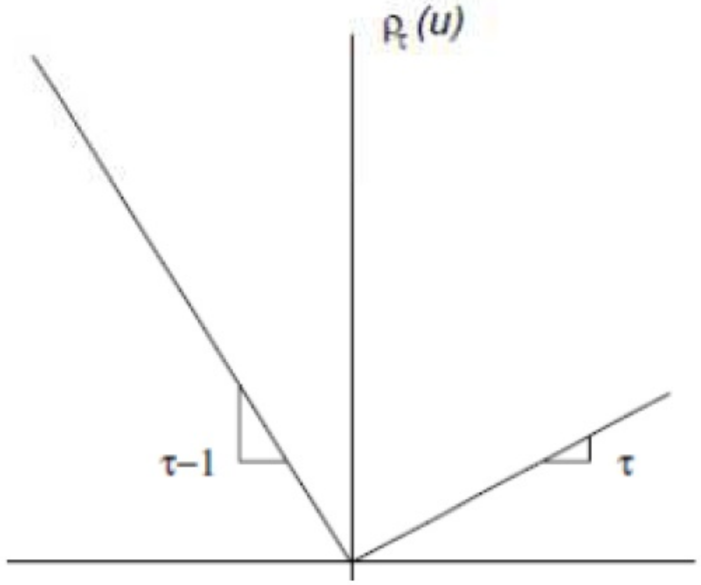

FIGURE 2.3.1. Quantile regression ρ function

To see this problem yields the sample quantiles as its solutions, it is only necessary to compute the directional derivative of the objective function with respect to ξ, taken from the left and from the right. Having succeeded in defining the unconditional quantiles as an optimization problem, it is easy to define conditional quantiles in an analogous fashion. Least squares regression offers a model for how to proceed. If, presented with a random sample $\{y_1, y_2, ..., y_n\}$,we solve

$$\min_{\mu \in \Re} \sum_{i=1}^{n} (y_i - \mu)^2, \tag{2.3.47}$$

and we proceed as seen previously.

In quantile regression, to obtain an estimate of the conditional median function, we simply replace the scalar ξ in 2.3.46 with the parametric function $\xi(x_i, \beta)$ and set τ to $\frac{1}{2}$.

To illustrate the basic ideas they reconsider a classical empirical application, Ernst Engel's (1875) analysis of the relationship between house food expenditure and household income. In Figure 2.3.2

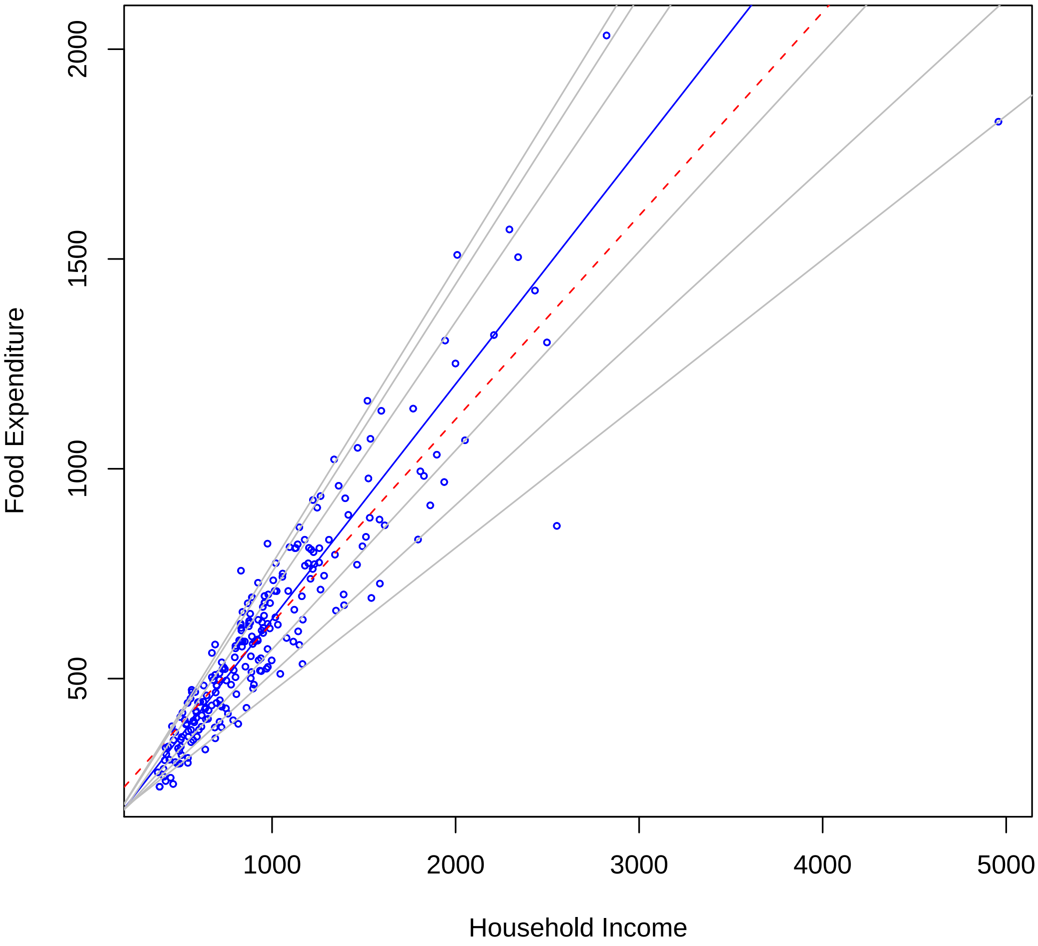

FIGURE 2.3.2. Engel Curves for Food

They plot the data taken from 235 European working class households. Superimposed on the plot are seven estimated quantile regression lines corresponding to the quantiles $\tau \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. The median $\tau = 0.5$ fit is indicated by dashed line; the least squares fit is plotted as the dotted line.

The plot reveals the tendency of the dispersion of food expenditure to increase along with its level as household income increases. The spacing of the quantile regression lines also reveals thet the conditional distribution of food expenditure is skewed to the left ; the narrower spacing of the upper quantiles indicating high density and a short upper tail and the wider spacing of the lower quantiles indicateing a lower density and longer lower tail. The conditional median and mean fits are quite different in this example, a fact that is partially explained by the asymmetry of the conditional density and partially by the strong effect exerted on the least squares fit by the two unusual points with high income and low food expenditure. Note that one consequence of this non robustness is that the least squares fit provides a rather poor estimate of the conditional mean for the poorest households in the sample; note that the dashed least squares line passes above all of the very low income observations.

EXAMPLE 2. **Quantile Regression and Determinants of Infant Birthweight**

Koenker and Hallock [2001], for this example, reconsider an investigation by Abreveya (2001) of the impact of various demographic characteristics and maternal behavior on the birthweight of in fants born in the United States. Low birthweight is know to be associated with a wide range of subsequent health problems, and has even been linked to educational attainment and eventual labor market outcomes; consequently , there has been considerable interest in factors influencing birthweights and public policy initiatives that might prove effective in reducing the incidence of low birthweight infants.

Their analysis is based on the June 1997 Detailed Natality Data published by the National Center for Health Statistics. Like Abreveya (2001) , they limit the sample to live, singleton births, with mother recorded as either black or white, between the age of 18 and 45, residing in the United States. Observations with missing data for any of the variables described below were dropped from the analysis. This process yielded a sample of 198.377 babies. Birthweight, the response, variable, is recorded in grams. Education of the mother is divided into four categories : less than high school, high school, some college and college graduate. The omitted category is less than high school so coefficients may be interpreted relative to this category. The prenatal medical care of the mother is also divided into four categories: those with no prenatal visit, those whose first prenatal visit was in the first trimester of the pregnancy, those with first visit in the second trimester, and those first visit in the last trimester. The omotted category is the group with a first visit in the first trimester, they constitute almost 85 percent of the sample. An indicator of whether the mother smoked during pregnancy is included in the model, as well as mother's reported average number of cigarettes smoked per day. The mother's reported weight gain durinf pregnancy (in pounds) is included (as a quadratic effect).

Figure 2.3.3 presents a summary of the quantile regression results for this example. They have 15 covariates, plus an intercept. For each of the 16 coefficients they plot the 19 distinct quantile regression estimates for τ ranging from 0.05 to 0.95 as the solid curve with filled dots. For each covariate these point estimates may be interpreted as the impact of a one unit change of the covariate on birthweight holding other covariates fixed. Thus, each of the plots have a horizontal quantile, or τ,scale and the vertical scale in grams indicates the covariate effect. The dashed line in each figure shows the ordinary least squares of the conditional mean effect. The two dotted lines represent conventional 90 percent confidence interval for the least squares estimate. The shaded grey area depicts a 90 percent pointwise confidence band for the quantile regression estimates.

In the first panel of the figure the intercept of the model may be interpreted as the estimated conditional quantile finction of the birthweight distribution of a girl born to an unmarried, white mother with less than a high school education, who is 27 years old and in the first trimester of the pregnancy. The mother's age and weight gain are chosen to reflect the means of these variables in the sample.

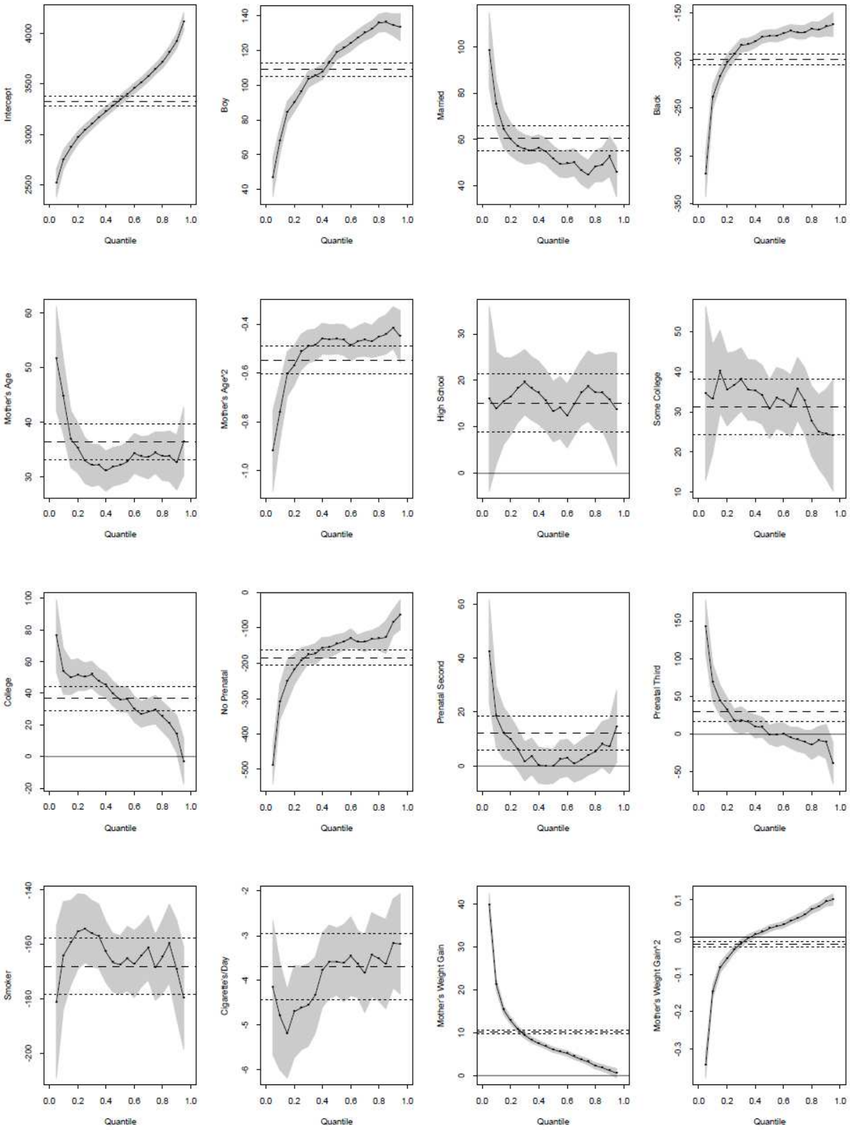

FIGURE 2.3.3. OLS and Quantile Regression Estimates for Birthweight Model

CHAPTER 3

Application

3.1. Introduction

The study of the data and the application of the presented methods is composed of two parts.

In the first part, all possible samples were extracted, in order to obtain the true distribution and thus to have different scenarios as much realistic as possible.

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.

In the second part, we tested all methods considered on a bivariate model. After having selected 1000 samples of size $n = 100$ from a Bivariate Log-Normal distribution, expected values and variances of the estimators are compared.

3.2. Examples of real distribution of the estimators

In order to extract all the possible samples, known 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

3.2.1. Median estimation without the auxiliary variable. The estimate of the median without the use of auxiliary variables is reduced to the calculation of thr median of the sample values.

Case 1 : X symmetric

Expected values

Y symmetric	Y positive asymmetric	Y negative asymmetric
35.1157	45.0435	55.8184

Variances and MSE

	Y symmetric	Y positive asymmetric	Y negative asymmetric
Var	140.2824	569.2835	1236.7778
MSE	140.2958	626.1874	1239.606

3.2.2. Kuk and Mak estimator [Kuk and Mak, 1989]. We consider all possible samples. For every sample we order the values of auxiliary variable X and variable Y in a non decreasing way.

If the value of X is less than or equal to the median Me_X, then add $\frac{1}{n}$ to the function $F_{Y_1}(y)$, if the value of X is greater than the median Me_X, then add $\frac{1}{n}$ to the function $F_{Y_2}(y)$.

Then $F_Y(y)$ is estimated by

$$\widetilde{F}_Y(y) \simeq \frac{1}{2}\left\{\tilde{F}_{Y_1}(y) + \tilde{F}_{Y_2}(y)\right\}$$

when the value $\frac{1}{2}$ is obtained, we'll get the corresponding median of Y.

Case 1 : X symmetric

	Y sym	**Y pos asym**	**Y neg asym**
Expected value	44.9923	68.0067	78.0784
Variance	121.862	1034.45	716.860
MSE	221.708	1965.1087	1140.3305

Case 2 : X positive asymmetric

	Y sym	**Y pos asym**	**Y neg asym**
Expected value	44.9923	68.0067	78.0784
Variance	121.862	1034.45	716.860
MSE	221.708	1965.1087	1140.3305

Case 3 : X negative asymmetric

	Y sym	**Y pos asym**	**Y neg asym**
Expected value	44.9923	68.0067	78.0784
Variance	121.862	1034.45	716.860
MSE	221.708	1965.1087	1140.3305

3.2.3. Ratio estimator. We apply the procedure of paragraph 2.3.2.

Case 1 : X symmetric

	Y sym	**Y pos asym**	**Y neg asym**

Expected value	35	43.588	50.3017
Variance	0	102.0516	506.1431
MSE	0	175.8047	557.9591

Case 2 : X poisitve asymmetric

	Y sym	**Y pos asym**	**Y neg asym**
Expected value	31.387	37.5	45.44975
Variance	32.7882	0	649.0173
MSE	45.8419	0	712.2159

Case 3 : X negative asymmetric

	Y sym	**Y pos asym**	**Y neg asym**
Expected value	49.2475	62.5704	57.5
Variance	467.9422	814.9926	0
MSE	670.9347	1575.119	0

3.2.4. Median regression. We apply the procedure of paragraph 2.3.3. We shall restrict this line to pass through the median values of these data.

A1. X symmetric and Y symmetric

First we calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{3+4}{2} \right\} = 3.5$$

$$Me(Y) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{30+40}{2} \right\} = 35.$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)
x_i	y_i	$x_i^{'}$	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$	Rank	$\tilde{x}_j$
1	10	-2.5	-25	10	1	-2.5
2	20	-1.5	-15	10	2	-1.5
3	30	-0.5	-5	10	3	-0.5
4	40	0.5	5	10	4	0.5
5	50	1.5	15	10	5	1.5
6	60	2.5	25	10	6	2.5

TABLE 11. Determining the slope of the line when X and Y are symmetric

(1) Columns (1), (2) are the recordings of raw data.

(2) Columns (3), (4) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using

$$x_i^{'} = x_i - Me(X)$$

$$y_i^{'} = y_i - Me(Y).$$

(3) Column (5) is the ratio $\frac{y_i^{'}}{x_i^{'}}$; these numbers are the b values at the minimum points of the indiviadual curves of the terms $\mid y_i^{'} - bx_i^{'} \mid$.

(4) Column (6) ranks the data of column (5) in ascending algebraic order.

(5) By adding the absolute values of column (3), we obtain $\sum \mid x_i^{'} \mid = 9$.

(6) The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until change in sign is obtained. Thus:

$$\begin{array}{rr} -\sum \mid x_i' \mid: & \text{-9} \\ 2\mid \tilde{x}_1 \mid: & \underline{\;5\;} \\ & \text{-4} \\ 2 \mid \tilde{x}_2 \mid: & \underline{\;3\;} \\ & \text{-1} \\ 2 \mid \tilde{x}_3 \mid: & \underline{\;1\;} \\ & 0 \end{array}$$

Since the index i=3 effects the change in sign, $b = y_3^{'}/x_3^{'}$=10 is the slope of the line of best fit,

$$\hat{y} = 10x.$$

A2. X symmetric and Y positive asymmetric

(1)	(2)	(3)	(4)	(5)	(6)
$x_i^{'}$	y_i	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$	Rank	$\tilde{x}_i$
-2.5	20	-17.5	7	3	-0.5
-1.5	25	-12.5	8.33	4	0.5
-0.5	35	-2.5	5	1	-2.5
0.5	40	2.5	5	2	-1.5
1.5	80	42.5	28.33	5	1.5
2.5	120	82.5	33	6	2.5

TABLE 14. Determining the slope of the line when X is symmetric and Y positive asymmetric

We calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{3+4}{2} \right\} = 3.5$$

$$Me(Y) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{35+40}{2} \right\} = 37.5.$$

(1) Columns (1), (3) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using

$$x_i^{'} = x_i - Me(X)$$

$$y_i^{'} = y_i - Me(Y).$$

(2) Column (4) is the ratio $\frac{y_i^{'}}{x_i^{'}}$; these numbers are the b values at the minimum points of the indiviadual curves of the terms $\mid y_i^{'} - bx_i^{'} \mid$.

(3) By adding the absolute values of column (1), we obtain $\sum \mid x_i^{'} \mid = 9$.

(4) The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until change in sign is obtained. Thus:

$$\begin{array}{rr} -\sum \mid x_i' \mid: & -9 \\ 2\mid \tilde{x}_1 \mid: & \underline{1} \\ & -8 \\ 2 \mid \tilde{x}_2 \mid: & \underline{1} \\ & -7 \\ 2 \mid \tilde{x}_3 \mid: & 5 \end{array}$$

$$\begin{array}{rc} & \overline{\text{-2}} \\ 2 \mid \tilde{x}_4 \mid: & \underline{3} \\ & 1 \end{array}$$

Since the index i=4 effects the change in sign, $b = y_2^{'}/x_2^{'}$=8.33 is the slope of the line of best fit,

$$\hat{y} = 8.33x.$$

A3. X symmetric and Y negative asymmetric

First we calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{3+4}{2} \right\} = 3.5$$

$$Me(Y) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{25+90}{2} \right\} = 57.5.$$

(1)	(2)	(3)	(4)	(5)	(6)
$x_i^{'}$	y_i	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$	Rank	$\tilde{x}_i$
-2.5	10	-47.5	19	2	2.5
-1.5	15	-42.5	28.33	4	-2.5
-0.5	25	-32.5	65	5	1.5
0.5	90	32.5	65	6	-1.5
1.5	91	33.5	22.33	3	-0.5
2.5	92	34.5	13.8	1	0.5

TABLE 17. Determining the slope of the line when X is symmetric and Y negative asymmetric

(1) Columns (1), (3) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using

$$x_i^{'} = x_i - Me(X)$$
$$y_i^{'} = y_i - Me(Y).$$

(2) Column (4) is the ratio $\frac{y_i^{'}}{x_i^{'}}$; these numbers correspondent to the b values at the minimum points of the indiviadual curves of the terms $\mid y_i^{'} - bx_i^{'} \mid$.

(3) By adding the absolute values of column (1), we obtain $\sum \mid x_i^{'} \mid = 9$.

(4) The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until change in sign is obtained. Thus:

$$\begin{array}{rr} -\sum \mid x_i' \mid: & \text{-9} \\ 2\mid \tilde{x}_1 \mid: & \underline{\;\;5\;\;} \\ & \text{-4} \\ 2 \mid \tilde{x}_2 \mid: & \underline{\;\;5\;\;} \\ & 1 \end{array}$$

Since the index i=2 effects the change in sign, $b = y_1^{'}/x_1^{'}$=19 is the slope of the line of best fit :

$$\hat{y} = 19x.$$

A4. X negative asymmetric and Y symmetric

(1) x_i	(2) y_i	(3) $x_i^{'}$	(4) $y_i^{'}$	(5) $b = \frac{y_i^{'}}{x_i^{'}}$	(6) Rank	(7) $\tilde{x}_i$
1	10	-4.75	-25	5.26	5	-3.25
1.5	20	-4.25	-15	3.53	3	3.25
2.5	30	-3.25	-5	1.54	1	-4.25
9	40	3.25	5	1.54	2	3.25
9.1	50	3.35	15	4.48	4	3.35
9.2	60	3.45	25	7.25	6	3.45

TABLE 20. Determining the slope of the line when X is negative asymmetic and Y symmetric

First we calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{2.5 + 9}{2} \right\} = 5.75$$

$$Me(Y) = \left\{ \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{30 + 40}{2} \right\} = 35.$$

(1) Columns (3), (4) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using

$$x_i^{'} = x_i - Me(X)$$

$$y_i^{'} = y_i - Me(Y).$$

(2) Column (5) is the ratio $\frac{y_i^{'}}{x_i^{'}}$; these numbers are the b values at the minimum points of the indiviadual curves of the terms $\mid y_i^{'} - bx_i^{'} \mid$.

(3) Column (6) ranks the data of column (5) in ascending algebraic order.

(4) By adding the absolute values of column (3), we obtain $\sum \mid x_i^{'} \mid = 20.53$.

(5) The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until change in sign is obtained. Thus:

$$
\begin{array}{rr}
-\sum \mid x_i' \mid: & \text{-20.53} \\
2\mid \tilde{x}_1 \mid: & 6.5 \\ \cline{2-2}
 & \text{-14.03} \\
2 \mid \tilde{x}_2 \mid: & 6.5 \\ \cline{2-2}
 & \text{-7.53} \\
2 \mid \tilde{x}_3 \mid: & 8.5 \\ \cline{2-2}
 & 1.03
\end{array}
$$

Since the index i=3 effects the change in sign, $b = y_2^{'}/x_2^{'}$=3.53 is the slope of the line of best fit,

$$\hat{y} = 3.53x.$$

A5. X negative asymmetric and Y positive asymmetric

First we calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{3+4}{2} \right\} = 5.75$$

$$Me(Y) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{+40}{2} \right\} = 37.5.$$

(1)	(2)	(3)	(4)	(5)	(6)
$x_i^{'}$	y_i	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$	Rank	$\tilde{x}_i$
-4.75	20	-17.5	3.68	4	-3.25
-4.25	25	-12.5	2.94	3	3.25
-3.25	35	-2.5	0.77	1	-4.25
3.25	40	2.5	0.77	2	-4.75
3.35	80	42.5	12.69	5	3.35
3.45	120	82.5	23.91	6	3.45

TABLE 23. Determining the slope of the line when X is negative asymmetric and Y positive asymmetric

(1) Columns (1), (3) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using

$$x_i^{'} = x_i - Me(X)$$

$$y_i^{'} = y_i - Me(Y).$$

(2) By adding the absolute values of column (1), we obtain $\sum \mid x_i^{'} \mid = 20.53$.

(3) The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until change in sign is obtained. Thus:

$$\begin{array}{rr} -\sum \mid x_i' \mid: & \text{-20.53} \\ 2| \tilde{x}_1 \mid: & \underline{6.5} \\ & \text{-14.03} \\ 2 \mid \tilde{x}_2 \mid: & \underline{6.5} \\ & \text{-7.53} \\ 2 \mid \tilde{x}_3 \mid: & \underline{12.75} \\ & 5.22 \end{array}$$

Since the index i=3 effects the change in sign, $b = y_2^{'}/x_2^{'}$=2.94 is the slope of the line of best fit,

$$\hat{y} = 2.94x.$$

A6. X negative asymmetric and Y negative asymmetric

We calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{2.5 + 9}{2} \right\} = 5.75$$

$$Me(Y) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{25 + 90}{2} \right\} = 57.5.$$

(1)	(2)	(3)	(4)
$x_i^{'}$	y_i	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$
-4.75	10	-47.5	10
-4.25	15	-42.5	10
-3.35	25	-32.5	10
3.25	90	32.5	10
3.35	91	33.5	10
3.45	92	34.5	10

Table 26. Determining the slope of the line when X is negative asymmetric and Y negative asymmetric

The slope of the line is $b = 10$.

A7. X positve asymmetric and Y symmetric

First we calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{3.5 + 4}{2} \right\} = 3.75$$

$$Me(Y) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{30 + 40}{2} \right\} = 35.$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)
x_i	y_i	$x_i^{'}$	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$	Rank	$\tilde{x}_i$
2	10	-1.75	-25	14.29	4	8.25
2.3	20	-1.45	-15	10.35	3	4.25
3.5	30	-0.25	-5	20	5	-1.45
4	40	0.25	5	20	6	-1.75
8	50	4.25	15	3.53	2	-0.25
12	60	8.25	25	3.03	1	0.25

Table 28. Determining the slope of the line when X is positive asymmetric and Y symmetric

(1) Columns (3), (4) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using

$$x_i^{'} = x_i - Me(X)$$

$$y_i^{'} = y_i - Me(Y).$$

(2) Column (5) is the ratio $\frac{y_i^{'}}{x_i^{'}}$; these numbers are the b values at the minimum points of the indiviadual curves of the terms $\mid y_i^{'} - bx_i^{'} \mid$.

(3) Column (6) ranks the data of column (5) in ascending algebraic order.

(4) By adding the absolute values of column (3), we obtain $\sum \mid x_i^{'} \mid = 16.2$.

(5) The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until change in sign is obtained. Thus:

$$\begin{array}{rr} -\sum \mid x_i' \mid: & \text{-16.2} \\ 2\mid \tilde{x}_1 \mid: & \underline{\quad 16.5} \\ & 0.3 \end{array}$$

Since the index i=1 effects the change in sign, $b = y_6^{'}/x_6^{'}$=3.03 is the slope of the line of best fit,

$$\hat{y} = 3.03x.$$

A8. X positive asymmetric and Y negative asymmetric

First we calculate the median of X and Y:

$$Me(X) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{3.5 + 4}{2} \right\} = 3.75$$

$$Me(Y) = \left\{ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \right\} = \left\{ \frac{x_{(3)} + x_{(4)}}{2} \right\} = \left\{ \frac{25 + 90}{2} \right\} = 57.5.$$

(1)	(2)	(3)	(4)	(5)	(6)
$x_i^{'}$	y	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$	Rank	$\tilde{x}_i$
-1.75	10	-47.5	27.14	3	8.25
-1.45	15	-42.5	29.31	4	4.25
-0.25	25	-32.5	130	5	-1.75
0.25	90	32.5	130	6	-1.45
4.25	91	33.5	7.88	2	-0.25
8.25	92	34.5	4.18	1	0.25

TABLE 31. Determining the slope of the line when X is positive asymmetric and Y symmetric

(1) Columns (1), (3) are the transformation of variables x_i, y_i to $x_i^{'}$, $y_i^{'}$ using

$$x_i^{'} = x_i - Me(X)$$

$$y_i^{'} = y_i - Me(Y).$$

(2) By adding the absolute values of column (1), we obtain $\sum \mid x_i^{'} \mid = -16.2$.

(3) The final step is to add successively to $-\sum \mid x_i^{'} \mid$, twice the individual $\mid \tilde{x}_i \mid$in the sequence of their indicated ranks until change in sign is obtained. Thus:

$$\begin{array}{rr} -\sum \mid x_i' \mid: & \text{-16.2} \\ 2\mid \tilde{x}_1 \mid: & \underline{16.5} \\ & 0.3 \end{array}$$

Since the index i=1 effects the change in sign, $b = y_6^{'}/x_6^{'}$=4.18 is the slope of the line of best fit,

$$\hat{y} = 4.18x.$$

A9. X positive asymmetric and Y positive asymmetric

Applying the procedure of the paragraph 2.3.3 , the slope of the line is $b = 10$.

(1)	(2)	(3)	(4)
$x_i^{'}$	y_i	$y_i^{'}$	$b = \frac{y_i^{'}}{x_i^{'}}$
-1.75	20	-17.5	10
-1.45	25	-12.5	10
-0.25	35	-2.5	10
0.25	40	2.5	10
4.25	80	42.5	10
8.25	120	82.5	10

Table 34. Determining the slope of the line when X is positive asymmetric and Y symmetric

The expected values, variances and MSE.

Case 1 : X symmetric

	Y symmetric	**Y positive asymmetric**	**Y negative positive**
Expected value	35	44.4004	55.5921
Variance	0	206.7387	349.1249
MSE	0	295.1064	352.7651

Case 2 : X positive asymmetric

	Y symmetric	**Y positive asymmetric**	**Y negative positive**
Expected value	32.7967	37.5	52.3161
Variance	39.1736	0	876.0749
MSE	44.0281	0	1095.591

Case 3 : X negative asymmetric

	Y symmetric	**Y positive asymmetric**	**Y negative positive**
Expected value	35.8111	45.5466	57.5
Variance	59.2385	308.3228	0
MSE	59.8963	419.5538	0

MSE of median regression.

We apply 2.3.31 to the data.

Case 1 : X symmetric

	(a)	(b)	(c)	(d)	(e)
Y sym.	139.2747	1.392747	13.92747	0	0
Y pos. asym	611.9496	1.368313	24.57176	264.6812	264.6812051
Y neg. asym	1242.512	1.351852	36.26826	352.3362	352.336692

Table 38. MSE : X symmetric

where :
$E(\hat{M}e_Y - Me)^2 = (a)$
$E(\hat{M}e_X - Me)^2 = (b)$
$Cov(\hat{M}e_Y, \hat{M}e_X) = (c)$
$E(\hat{M}e_{MR} - Me_Y)^2 = (d)$
$2.3.31 = (e)$

Case 2 : X positive asymmetric

	(a)	(b)	(c)	$((d)$	(e)
Y symmetric	138,9918	6.12234	24,72209	45,38448	45,3845
Y pos. asymmetric	633,3122	6.33122	63.33122	0	0
Y neg asymmetric	1248,581	6.046307	53,30765	908.5728	908,5725404

Table 40. MSE : X positive asymmetric

Case 3 : X negative asymmetric

	(a)	(b)	(c)	(d)	(e)
Y symmetric	137,6029	12,44971	36,53646	60,1068	60.107
Y pos. asymmetric	656,7837	12,46848	51,93538	443,3925	443,3924468
Y neg. asymmetric	1252,531	12,52531	125,2531	0	0

Table 42. MSE : X negative asymmetric

3.2.5. *Linear regression.* We apply the procedure of paragraph 2.3.4.

Case 1 : X symmetric

	Y symmetric	**Y positive asymmetric**	**Y negative positive**
Expected value	35	44.7901	55.64535
Variance	0	120.1250	331.1316
MSE	0	173.2708	334.5713

Case 2 : X positive asymmetric

	Y symmetric	**Y positive asymmetric**	**Y negative positive**
Expected value	31.7568	37.5	49.44801
Variance	29.1739	0	714.3191
MSE	39.6922	0	779.1537

Case 3 : X negative asymmetric

	Y symmetric	**Y positive asymmetric**	**Y negative positive**
Expected value	35.868	46.6717	57.5
Variance	48.6476	416.5605	0
MSE	49.4010	500.6814	0

3.2.6. Quantile regression.

We apply the procedure of paragraph 2.3.5 , when $\tau = \frac{1}{2}$.

Case 1 : X symmetric

	Y symmetric	Y positive asymmetric	Y negative positive
Expected value	35	44.7615	55.2354
Variance	0	123.2197	410.88
MSE	0	218.5056	414.793

Case 2 : X positive asymmetric

	Y symmetric	Y positive asymmetric	Y negative positive
Expected value	32.425	37.5	49.5849
Variance	33.4249	0	718.779
MSE	40.2642	0	781.428

Case 3 : X negative asymmetric

	Y symmetric	Y positive asymmetric	Y negative positive
Expected value	35.8178	46.6555	57.5
Variance	42.1649	411.3568	0
MSE	42.8337	495.1804	0

3.3. Final results

We compare the different estimators for the median.

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
Kuk and Mak	44.9923	68.0067	78.0784

Table 50. Expected values : X symmetric

Table 50 shows that

-for **Y symmetric:** the estimators are unbiased, except for the estimator of Kuk and Mak, that is biased;

-for **Y positive and negative asymmetric :** the estimators are biased and we note the high value for the estimator of Kuk and Mak.

To choose the best estimator, we observe the Tables 52 and 54 :

-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. The estimator of Kuk and Mak worses the median estimation with respect to the estimation without the auxiliary variable;

- 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
Kuk and Mak	121.862	1034.45	716.860

Table 52. 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
Kuk and Mak	221.708	1965.078	1140.2705

Table 54. MSE : X symmetric

Case1: 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
Kuk and Mak	44.9923	68.0067	78.0784

Table 56. Expected values : X positive asymmetric

Table 56 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, except for the estimator of Kuk and Mak.

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
Kuk and Mak	121.862	1034.45	716.860

Table 58. 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
Kuk and Mak	221.708	1965.078	1140.2705

Table 60. MSE : X positive asymmetric

Table 58 and 60 show that

- for **Y symmetric** and **Y positive asymmetric :** the best estimator is the linear regression's 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
Kuk and Mak	44.9923	68.0067	78.0784

Table 62. Expected values : X negative asymmetric

Table 62 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, except for estimators without auxiliary variable and Kuk and Mak.

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
Kuk and Mak	121.862	1034.45	716.860

Table 64. 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
Kuk and Mak	221.708	1965.078	1140.2705

Table 66. MSE : X negative asymmetric

Tables 64 and 66 show that:

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

3.4. Simulation : Bivariate Log-Normal distribution

The second application is related to a simulation involving a Bivariate Log-Normal distribution.

Here we describe the procedure.

We start by simulating a Bivariate Normal in R

$$rbinorm < -function(n, mux, muy, sigmax, sigmay, rho)$$

where n is the sample size, mux is the expected value of X, muy is the expected value of Y, $sigmax$ represents the standard deviation of X, $sigmay$ is the standard deviation of Y and rho is the linear correlation coefficient;

generate the marginal X from the the Normal distribution

$$x < -rnorm(n, mux, sigmax),$$

generate y from a conditional function $Y \mid X = x$

$$y < -rnorm(n.muy + rho * sigmay * (x - mux)/sigmax, sigmay * sqrt(1 - rho^2)\}.$$

To obtain the Bivariate Log-Normal distribution, we apply the exponential transformation of each component of a Bivariate Normal pair.

For the present experiment, the values assigned to the parameters are :

- n=1000,
- mux=1,
- muy=2,
- sx= 0.5,
- sy = 0.7

-rho= (0.4 ; 0.7; 0.9).

Chosen $\rho = 0.4$ and $Me(Y) = 7.21$, the results are

	Expected value	**Variance**	**MSE**
No auxiliary variable	7.2189	0.4032	0.4033
Ratio	7.2146	0.4221	0.4221
Median regression	7.209	0.378	0.3787
Linear regression	7.209	0.398	0.398

[Chu, 1955] shows that if the parent population is normal, then the distribution of the sample median tends "rapidly" to normality.

In fact the value of $Var(\hat{M}e(Y)) = 0.4032$ approximates well the variance proposed by Chu [1955]

$$\sigma_n^2 = \frac{1}{4\left[f(\xi)\right]^2(2n+1)} = 0.4019.$$

Similarly for the variable X, in fact $Var(\hat{M}e(X)) = 0.022$ approximates well the variance

$$\sigma_n^2 = \frac{1}{4\left[f(\xi)\right]^2(2n+1)} \simeq 0.029.$$

Chosen $\rho = 0.7$ and $Me(Y) = 7.29$, the results obtained through the experiment are

	Expected value	**Variance**	**MSE**
No auxiliary variable	7.2923	0.42	0.42
Ratio	7.2277	0.341	0.345
Median regression	7.2233	0.333	0.338
Linear regression	7.22	0.387	0.3919

The variance $Var(\hat{M}e_Y) = 0.42$ approximates well the variance proposed by Chu [1955]

$$\sigma_n^2 = \frac{1}{4\left[f(\xi)\right]^2(2n+1)} = 0.4011.$$

Similarly for the variable X, in fact $Var(\hat{M}e(X)) = 0.0288$ approximates the variance

$$\sigma_n^2 = \frac{1}{4\left[f(\xi)\right]^2(2n+1)} \simeq 0.029.$$

Chosen $\rho = 0.9$ and $Me(Y) = 7.69$, the results obtained are

	Expected value	**Variance**	**MSE**
No auxiliary variable	7.7	0.488	0.489
Ratio	7.715	0.2496	0.25
Median regression	7.72	0.2476	0.248
Linear regression	7.72	0.2486	0.2495

From the above experiment we obtain $Var(\hat{Me}(Y)) = 0.408$. This result approximates well the variance proposed by Chu (1955)

$$\sigma_n^2 = \frac{1}{4\left[f(\xi)\right]^2(2n+1)} \simeq 0.38.$$

It is possible to consider that the variances of all the estimators decrease when the correlation coefficient increases and the three estimators cosidered are almost unbiased.

The median regression estimator in all the three situations is the best.

Conclusions

Thanks to this research work of thesis 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.

- The estimator of Kuk and Mak doesn't improve with respect to the estimator of median without auxiliary variable; but if the mean square error of the present estimator is compared with the estimated mean square error of the other estimators, it gets worse. May be it depends on the fact tahat this method takes into consideration only the relative frequency od cases in which $X < Me_X$ and $Y < Me_Y$, but it doesn't take into consideration the real value of the median of the auxiliary variable.

- 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, as it has been observed in the last part of Chapter 3, 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.

We perfectly know that this topic deserves a wider and fuller treatment.

However, considering the many research ideas, that have been presented in the course of our study, we believe that this work can be used as a starting point for this kind of research, that coul reach interesting.

Bibliography

V. Barnett. Sample survey. *Principles and methods. London etc.: Edward Arnold*, 1991.

E. Brentari. Asimmetria e misure di asimmetria, 1990.

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

J.G. Eisenhauer. Symmetric or skewed? *The College Mathematics Journal*, 33 (1):48–51, 2002.

B.V. Frosini. *Lezioni di statistica.* Vita e Pensiero, 1990.

R.A. Groeneveld and G. Meeden. Measuring skewness and kurtosis. *The Statistician*, pages 391–399, 1984.

ST Gross. Median estimation in sample surveys. *Proceedings of the Survey Research Section*, pages 181–184, 1980.

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

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

R. Koenker and K.F. Hallock. Quantile regression. *The Journal of Economic Perspectives*, 15(4):143–156, 2001.

A.Y.C. Kuk and TK Mak. Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 261–269, 1989.

Angiola Pollastri. *Elementi di teoria dei campioni.* Edizioni CUSL, 1997.

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.

P. Radaelli. La regressione mediana. 2002.

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

C.E. Särndal, B. Swensson, and J. Wretman. *Model assisted survey sampling.* Springer Verlag, 2003.

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

R.S. Woodruff. Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47(260):635–646, 1952.

M. Zenga. *Introduzione alla statistica descrittiva.* Vita e pensiero, 1988.

Part 1

Appendix A

[Kuk and Mak, 1989]Let $F_Y(\cdot)$ be the cumulative distribution function of the density f_Y. Taylor's series expansion yields

$$F_Y(\hat{M}e_Y) = F_Y(Me_Y) + f_Y(Me_Y)(\hat{M}e_Y - Me_Y) + o_p(n^{-1/2}),$$

so that $\hat{M}e_Y - Me_Y = \{f_Y(Me_Y)\}^{-1}\left\{F_Y(\hat{M}e_Y) - F_Y(Me_Y)\right\} + o_p(n^{-1/2})$. Furthermore, it can be shown that

$$F_Y(\hat{M}e_Y) - F_Y(Me_Y) = \hat{F}_Y(\hat{M}e_Y) - \hat{F}_Y(Me_Y) + o_p(n^{-1/2}).$$

Therefore

$$\begin{aligned}\hat{M}e_Y - Me_Y = &\quad \{f_Y(Me_Y)\}^{-1}\left\{\hat{F}_Y(\hat{M}e_Y) - \hat{F}_Y(Me_Y)\right\} + o_p(n^{-1/2}) \\ = &\quad \{f_Y(Me_Y)\}^{-1}\left(\tfrac{1}{2} - p_Y\right) + o_p(n^{-1/2}),\end{aligned} \tag{.0.1}$$

where $p_Y = \hat{F}_Y(Me_Y)$. With $p_X = \hat{F}_X(Me_X)$, we also have

$$\hat{M}e_X - Me_X = \{f_X(Me_X)\}^{-1}\left(\frac{1}{2} - p_X\right) + o_p(n^{-1/2}). \tag{.0.2}$$

It follows from equation .0.1 and .0.2 that the asymptotic distribution of $(\hat{M}e_Y - Me_Y, \hat{M}e_X - Me_X)$ is Bivariate Normal with means zero, variances

$$\sigma_Y^2 = (1-f)(4n)^{-1}\{f_Y(Me_Y)\}^{-2},$$

$$\sigma_X^2 = (1-f)(4n)^{-1}\{f_X(Me_X)\}^{-2}$$

and covariance

$$\sigma_{XY} = (1-f)n^{-1}\left(P_{11} - \frac{1}{4}\right)\{f_X(Me_X)f_Y(Me_Y)\}^{-1}.$$

Kuk and Mak [1989] now derive the asymptotic distribution of $\hat{M}e_{YR} - Me_Y = (Me_X\hat{M}e_Y - Me_Y\hat{M}e_X)/\hat{M}e_X$.

Since $\hat{M}e_X/Me_X \to 1$ in probability, $\hat{M}e_{YR} - Me_Y$ has the same asymptotic distribution as

$$(Me_X\hat{M}e_Y - Me_Y\hat{M}e_X)/Me_X = (\hat{M}_Y - M_Y) - (M_Y - M_X)(\hat{M}_X - M_X).$$

Thus $\hat{M}e_{YR}$is asymptotically normal with mean $Me_Y$and variance

$$\begin{gathered} n^{-1}(1-f)[\tfrac{1}{4}\{f_Y(M_Y)\}^{-2} + \tfrac{1}{4}(M_Y/M_X)^2\{f_X(M_X)\}^{-2} \quad + \\ -2(M_Y/M_X)\left\{f_Y(M_Y)f_X(M_X)\right\}^{-1}(P_{11} - \tfrac{1}{4})]. \end{gathered}$$

They next derive the asymptotic distribution of $\hat{M}_{YP} = \hat{Q}_Y(\hat{p}_1)$.

Now it is easy to show that

$$\hat{p}_1 - \hat{p}_0 = 2(p_{11} - P_{11})(2p_X - 1).$$

Consequently $(\hat{p}_1 - \hat{p}_0)\sqrt{n} \rightarrow 0$ in probability since $p_X \rightarrow \frac{1}{2}$ in probability and $p_{11} - P_{11}$is of order $O_p(n^{-1/2})$.

Thus

$$\hat{M}e_{YP} - Me_Y = \{f_Y(Me_Y)\}^{-1}(\hat{p}_0 - p_Y) + o_p(n^{-1/2}).$$

Hence, to find the asymptotic variance of $\hat{M}e_{YP}$ it suffices to find $E(\hat{p}_0 - p_Y)^2$. Some direct algebraic manupulation yields

$$\hat{p}_0 - p_Y = (4P_{11} - 1)(p_X - \frac{1}{2}) - (p_Y - \frac{1}{2})$$

so that

$$E(\hat{p}_0 - p_Y)^2 = 2(1-f)P_{11}(1-2P_{11})n^{-1}.$$

Hence, $\hat{M}e_{YP}$ is asymptotically normal with mean Me_Y and variance

$$2\{f_Y(Me_Y)\}^{-2}(1-f)P_{11}(1-2P_{11})n^{-1} \tag{.0.3}$$

To derive the asymptotic variance of $\hat{M}e_{YS,}$it is similar noted that:

$$\hat{M}e_{YS} - Me_Y = \{f_Y(Me_Y)\}^{-1}\left\{\frac{1}{2} - \tilde{F}(Me_Y)\right\} + o_p(n^{-1/2}).$$

Now $\tilde{F}(Me_Y) = \frac{1}{2}\left\{\tilde{F_{Y1}}(Me_Y) + \tilde{F}_{Y2}(Me_Y)\right\}$and hence, conditional on n_X,

$E\left\{\tilde{F}(Me_Y)\right\} = \frac{1}{2}(2P_{11} + 2P_{12})$ and

$$var\left\{\tilde{F}(Me_Y)\right\} = \frac{1}{4}\left\{(1-f_1)\,2P_{11}(1-2P_{11})n_X^{-1} + (1-f_2)2P_{12}(1-2P_{12})(n-n_X)^{-1}\right\},$$

where $f_1 = 2n_X/N$ and $f_2 = 2(n-n_X)/N$. Thus for large n unconditional variance of $\tilde{F}(M_Y)$ is

$$(1-f)\{P_{11}(1-2P_{11}) + P_{12}(1-2P_{12})\}n^{-1} =$$

$$= 2(1-f)P_{11}(1-2P_{11})n^{-1}.$$

It follows that expression .0.3 is also the asymptotic variance of $\hat{M}e_{YS}$.