International Journal of Business and Statistical Analysis
ISSN (2384-4663)
Int. J. Bus. Stat. Ana. 2, No. 1 (Jan-2015)

New Formulae and Uses of Mean Absolute Deviation

Elsayed A. H. Elamir

Department of Statistics and Mathematics,Faculty of Commerce, Benha University, Egypt
Management & Marketing Department, College of Business, University of Bahrain

Received 1 December 2014, Revised 15 December 2014, Accepted 16 December 2014, Published 1 January 2015

Abstract: Mean absolute deviation about mean and median is extended to define "MAD-comoment" similarly analogous to classical central moment notation of covariance. Therefore, the "MAD-comoment" is used to derive the coefficient of determination in LAD regression, to analyze total sum of absolute into the pure between sum of absolute and the pure within sum of absolute that provides the exact decomposition of Pietria index of inequality, to develop a new formula for beta coefficient in financial risk analysis that gives less weight at the extreme ends than least squares method for market return and to define MAD-correlation under first moment assumption.

Keywords: beta coefficient; comoment; correlation; MAD; LAD regression; Pietra index.

1 Introduction

Mean absolute deviation (MAD) offers a direct and robust measure of the dispersion of a random variable and has many applications in different fields; see, Bassett and Koenker (1978), Dodge (1987), Bloomfield and Steigler (1983), Pham-Gia and Hung (2001), Gorard (2005), Habib (2011) and Elamir (2012). The mean absolute deviation is actually more efficient than standard deviation in life-like situations where small errors will occur in observations and measurements; see, Tukey (1960) and Huber (1981).

The univariate MAD about mean and median is extended to MAD-comoment, similarly analogous to classical central moment notions of covariance. The "MAD-comoment" introduces an easy and simple way to obtain the coefficient of determination in LAD regression, to analyze total sum of absolute into the pure between sum of absolute and the pure within sum of absolute that provides the exact decomposition of Pietra index of inequality, to develop a beta coefficient in financial risk analysis that gives less weights for market return at extreme ends than the least square beta coefficient and to yield correlation coefficient not only coherent with classical correlation, but also valid and meaningful under just first moment assumption.

In Section 2 the representation of MAD in terms of covariance is presented. In Section 3 the population MAD-comoment analog to central moment notions and their sample counterparts are introduced. Applications of MAD-comoment to LAD regression for finding the coefficient of determination, partitions total sum of absolute into pure between and within sum of absolute, developing the beta coefficient in financial risk analysis and finding a correlation coefficient based on MAD are studied in Section 4. Section 5 is devoted to the conclusion.

2 Representation of MAD as a covariance

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a continuous distribution with, density function $f(y)$, quantile function $y(F) = F^{-1}(y) = Q(F)$, $0 < F < 1$, cumulative distribution function $F(y) = F$, mean $\mu = E(Y)$ and median $\nu = Med(Y)$. The population MAD about mean and median is defined as

$$D_\mu = E|Y - \mu| \quad \text{and} \quad D_\nu = E|Y - \nu| \qquad (1)$$

Let over indicator functions are

$$I_O(Y,\mu) = I_{O\mu} = \begin{cases} 1, & Y > \mu \\ 0, & \text{else} \end{cases} \quad \text{and} \quad I_O(Y,\nu) = I_{O\nu} = \begin{cases} 1, & Y > \nu \\ 0, & \text{else} \end{cases} \qquad (2)$$

Habib (2011) and Elamir (2012) used the general dispersion function that defined by Munoz-Perez and Sanchez-Gomez (1990) as

E-mail: shabib40@gmail.com

$$D_Y(a) = E|Y-a| = a[2F_y(a)-1] + E(Y) - 2\int YI_{Y<a}dP \quad (3)$$

to re-define the population MAD about mean and median in terms of over indicator functions as

$$D_Y(\mu) = D_\mu = E|Y-\mu| = E\{2[I_{0\mu} - E(I_{O\mu})]Y\}, \quad (4)$$

$$D_Y(\nu) = D_\nu = E|Y-\nu| = E[(2I_{O\nu}-1)Y]$$

Consequently, the population MAD is re-defined as twice the covariance between the random variable Y and its indicator function as

$$D_\mu = 2[E(YI_{O\mu}) - E(Y)E(I_{O\mu})] = 2Cov(Y, I_{O\mu}),$$
$$D_\nu = 2E(YI_{O\nu}) - 2E(Y)E(I_{O\nu}) = 2Cov(Y, I_{O\nu}) \quad (5)$$

Note that $E(I_{o\mu}) = F(\mu)$, $V(I_{o\mu}) = F(\mu)(1-F(\mu))$, $E(I_{ov}) = 0.5$ and $V(I_{ov}) = 0.25$.

3 MAD-comoment

Following Serfling and Xiao (2007) consider a bivariate random variable $(Y^{(1)}, Y^{(2)})$ having cdf F with marginal distributions F_1 and F_2 means $\mu_1^{(1)}$ and $\mu_1^{(2)}$, medians $\nu^{(1)}$, and $\nu^{(2)}$ and finite central moments $\mu_k^{(1)}$ and $\mu_k^{(2)}, k = 1,2, \ldots$, the kth central comoment of $Y^{(1)}$ with respect to $Y^{(2)}$ is defined as

$$\xi_{k[12]} = cov\left(Y^{(1)}, \left(Y^{(2)} - \mu_1^{(2)}\right)^{k-1}\right) \quad (6)$$

of course for the second order case we have simply $\xi_{2[12]} = \sigma_{12}$ the usual covariance. The usual correlation coefficient denoted $\rho_{12} = \xi_{2[12]}/\sigma_1\sigma_2$.

3.1 Population MAD-comoment

Using the covariance representation (5) for MAD, and also by analogy with the central comoment given in (6) at $k = 2$, the MAD-comoment in terms of mean is defined as

$$D_{\mu[12]}(Y^{(1)}, Y^{(2)}) = Cov\left(Y^{(1)}, 2I_{O\mu}^{(2)} - 2F(\mu_1^{(2)})\right),$$
$$D_{\mu[21]}(Y^{(2)}, Y^{(1)}) = Cov\left(Y^{(2)}, 2I_{O\mu}^{(1)} - 2F(\mu_1^{(1)})\right) \quad (7)$$

Here $D_{\mu[12]}$ and $D_{\mu[21]}$ need not be equal. Similarly in terms of median

$$D_{\nu[12]} = Cov(Y^{(1)}, 2I_{O\nu}^{(2)} - 1),$$
$$D_{\nu[21]} = Cov(Y^{(2)}, 2I_{O\nu}^{(1)} - 1) \quad (8)$$

These are termed as MAD-comoment of $Y^{(1)}$ with respect to $Y^{(2)}$ and MAD-comoment of $Y^{(2)}$ with respect to $Y^{(1)}$. It is readily seen that MAD-comoment is a translation invariant and scale equivariant, that is,

$$D_{[12]}(a + bY^{(1)}, c + dY^{(2)}) = bD_{[12]}(Y^{(1)}, Y^{(2)}) \quad (9)$$

For positive b and d and arbitrary a and c.
For $Y^{(1)} = Y^{(2)}$, the MAD-comoment reduce to the MAD. On the other hand for $Y^{(1)}$ and $Y^{(2)}$ are independent the MAD-comoment takes the value zero.

3.2 Sample MAD-comoment

Consider a bivariate data pairs $(y_i^{(1)}, y_i^{(2)}), i = 1, \ldots, n$. From (7) the direct estimate of MAD-comoment using mean is

$$d_{\bar{y}[12]} = \frac{(n-1)}{n} Cov(y^{(1)}, 2I_{O\bar{y}}^{(2)}) \text{ and } d_{\bar{y}[21]} = \frac{(n-1)}{n} Cov(y^{(2)}, 2I_{O\bar{y}}^{(1)}) \quad (10)$$

From (8) the direct estimate of MAD-comoment using median is

$$d_{\tilde{y}[12]} = \frac{(n-1)}{n} Cov(y^{(1)}, 2I_{O\tilde{y}}^{(2)}) \text{ and } d_{\tilde{y}[21]} = \frac{(n-1)}{n} Cov(y^{(2)}, 2I_{O\tilde{y}}^{(1)}) \quad (11)$$

In general $d_{[12]}$ and $d_{[21]}$ are not equal, to obtain equal MAD-comoment, the central MAD-comoment may be used

$$h_1 = \frac{d_{\bar{y}[12]} + d_{\bar{y}[21]}}{2} \text{ and } h_2 = \frac{d_{\tilde{y}[12]} + d_{\tilde{y}[21]}}{2} \quad (12)$$

and

$$I_{\bar{y}}^{(1)} = \begin{cases} 1, & y_{(i)}^{(1)} > \bar{y}^{(1)} \\ 0, & \text{else} \end{cases} \text{ and } I_{\bar{y}}^{(2)} = \begin{cases} 1, & y_{(i)}^{(2)} > \bar{y}^{(2)} \\ 0, & \text{else} \end{cases} \quad (13)$$

Simiraly $I_{\tilde{y}}^{(1)}$ and $I_{\tilde{y}}^{(2)}$ can be defined by replacing the sample median rather than sample mean in (13).
Another useful form for MAD-comoment in terms of weighted average can be obtained as

$$d_{[12]} = \sum_{i=1}^{n} w_i y_i^{(1)} \quad \text{and} \quad d_{[21]} = \sum_{i=1}^{n} w_i y_i^{(2)} \tag{14}$$

where $w_i = \frac{2(n-1)}{n^2}(I_o - \bar{I}_o)$.

4 Applications of MAD-comoment

MAD-comoment has an important property for univariate $Y = Y_1 + \cdots + Y_n$ then

$$Cov(Y_1, \ldots, Y_M, I_O(Y)) = \sum_{m=1}^{M} Cov(Y_m, I_O(Y)) \tag{15}$$

This is called a semi decomposition of MAD-comoment and can be used in obtaining the coefficient of determination in LAD regression.

4.1 Coefficient of determination based on MAD-comoment

A nature measure in the regression of the effect of X in reducing the variation in Y, i.e., in reducing the uncertainty in predicting Y, is a measure of determination R^2. For least absolute deviation (LAD) regression

$$Min. |y_i - \acute{x}\beta|$$

The coefficient of determination of LAD regression can easily be obtained using MAD-comoment as follows.

Theorem 1

The coefficient of determination in LAD regression using MAD-comoment is

$$R^2_{MAD} = \frac{2Cov(\hat{y}_i, I_o(y))}{D_y} \tag{16}$$

Proof:

Since the actual value in LAD regression can be represented by

$$y_i = \hat{y}_i + e_i$$

By using (15) therefore

$$nD_y = 2nCov(y, I_O(y)) = 2nCov(\hat{y}_i, I_O(y)) + 2nCov(e_i, I_O(y))$$

This can be rewritten as

$$TAD = RAD + EAD \tag{17}$$

$TAD = 2nCov(y, I_O(y))$ is the total absolute deviation(TAD), $2nCov(\hat{y}_i, I_O(y))$ is the regression absolute deviation (RAD) and $2nCov(e_i, I_O(y))$ is the error absolute deviation (EAD).

By dividing both sides by nD_y

$$1 = \frac{2Cov(\hat{y}_i, I_O(y))}{D_y} + \frac{2Cov(e_i, I_O(y))}{D_y}$$

The explained variation (coefficient of determination) based on LAD regression line is

$$R^2_{MAD} = \frac{2Cov(\hat{y}_i, I_O(y))}{D_y} \tag{18}$$

The non-explained variation based on LAD regression line is

$$1 - R^2_{MAD} = \frac{2Cov(e_i, I_O(y))}{D_y} \tag{19}$$

4.1.1 Example

The data set given in Table 1 consists of a logarithmic deflated growth rate (Y) of a Finish company, a time variable (X_1), and a measure of the cyclical fluctuation in the aggregated funds from operations in logarithmic terms (X_2); see, Pynnonen and Salmi (1994).

Table 1. funds of operations and whole sale price index in years 1969-1982.

LNQ (Y)	Year (X_1)	LK3RES (X_2)	$\hat{Y}$	e	
11.654	1969	0.0434	11.754	-0.100	LAD Regression
12.099	1970	0.0911	11.794	0.304	$\hat{Y} = -178.01 + 0.096X_1 - 1.174X_2$
11.728	1971	0.0212	11.973	-0.245	
12.026	1972	0.0585	12.026	0	Semi-decomposition
12.252	1973	0.1020	12.071	0.181	$TAD = 6.46, RAD = 4.17, EAD = 2.29$
12.793	1974	0.0934	12.177	0.615	
13.163	1975	-0.0931	12.493	0.669	MAD determination (explained variation)
12.078	1976	-0.1196	12.620	-0.542	$R^2_{MAD} = 4.17/6.46 = 0.64$
12.805	1977	-0.1945	12.805	0	Non-explained variation
13.043	1978	-0.0620	12.745	0.297	$1 - R^2_{MAD} = 0.36$
12.094	1979	0.1383	12.607	-0.513	
12.521	1980	0.0776	12.774	-0.253	
12.917	1981	0.0386	12.917	0	
13.156	1982	-0.0407	13.106	0.050l	

The LAD regression is computed using package "quantreg" in R-sofware and R^2_{MAD} using MAD about median. Note that Pynnonen and Salmi (1994) used the measure

$$G = 1 - \sum_{i=1}^{n} \left| y_i - \acute{x}_i \tilde{\beta} \right| / \sum_{i=1}^{n} \left| y_i - med_y \right|$$

as a measure of determination that is equal to 0.417.

4.2 MAD partitions about mean and median

Assume there are G different groups with individuals in each group y_{gi}, $i = 1,2,\dots,n_g$, $n = n_1 + \dots + n_G$ and $g = 1,\dots,G$. Let $y_{gi} - \bar{y}$ is the total deviation ($\bar{y} = \sum_g^G \sum_i^{n_g} y_{gi}/n$), $\bar{y}_g - \bar{y}$ is the deviation of grouped mean ($\bar{y}_g = \sum_{i=1}^{n_g} y_{gi}/n_g$) around total mean, and $y_{gi} - \bar{y}_g$ is the deviation of individuals around the grouped mean.
By squaring and taking the summation over both g and i then

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{gi} - \bar{y})^2 = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\bar{y}_g - \bar{y})^2 + \sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{gi} - \bar{y}_g)^2$$

That is known as the partitions the total sum of squares, in other words, the total sum of squares (SST) is equal to the between sum of squares (SSB) plus the within sum of square (SSW), therefore

$$SST = SSB + SSW$$

The absolute value could be used rather than the square to give us

$$|y_{gi} - \bar{y}| \le |\bar{y}_g - \bar{y}| + |y_{gi} - \bar{y}_g|$$

By taking the summation over both g and i then

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \bar{y}| \le \sum_{g=1}^{G}\sum_{i=1}^{n_g}|\bar{y}_g - \bar{y}| + \sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \bar{y}_g|$$

The total sum of absolute ($TSA = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|x_{gi} - \bar{x}|$) is less than or equal to the between sum of absolute ($BSA = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|\bar{x}_g - \bar{x}|$) plus the within sum of absolute ($WSA = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|x_{gi} - \bar{x}_g|$), therefore,

$$TSA \le BSA + WSA$$

This can be re-written as

$$TSA = BSA + WSA + R$$

where $R = TSA - BSA - WSA$ is the residuals and it could be separated to pure within-groups residuals and pure between-groups residuals and added to each term to obtain perfect MAD partitions

$$TSA = (BSA - R_b) + (WSA - R_w) = PBSA + PWSA$$

This is shown in the following theorem.

Theorem 2
The MAD partitions about mean into the pure between sum of absolute and the pure within sum of absolute are

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \bar{y}| = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|\bar{y}_g - \bar{y}|_\delta + \sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \bar{y}_g|_\delta \qquad (20)$$

where

$$|\bar{y}_g - \bar{y}|_\delta = \begin{cases} -|\bar{y}_g - \bar{y}| & \text{if } y_{gi} \le \bar{y} < \bar{y}_g \text{ or } y_{gi} > \bar{y} \ge \bar{y}_g \\ |\bar{y}_g - \bar{y}| & \text{else} \end{cases}$$

and

$$|y_{gi} - \bar{y}_g|_\delta = \begin{cases} -|y_{gi} - \bar{y}_g| & \text{if } \bar{y}_g < y_{gi} \le \bar{y} \text{ or } \bar{y}_g \ge y_{gi} > \bar{y} \\ |y_{gi} - \bar{y}_g| & \text{else} \end{cases}$$

Note that $|\bar{y}_g - \bar{y}|_\delta$ is the pure between sum of absolute and $|y_{gi} - \bar{y}_g|_\delta$ is the pure within sum of absolute, i.e.

$$TSA = PBSA + PWSA$$

Proof:
Using the covariance representation of the mean absolute deviation about mean given in (5), the partitions can be written as $TSA = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \bar{y}| = 2Cov(y_{gi}, I_{o\bar{y}}) = 2E[(y_{gi} - \bar{y})I_{o\bar{y}}] = 2\sum_{g=1}^{G}\sum_{i=1}^{n_g} t_{gi}$

$$t_{gi} = \begin{cases} y_{gi} - \bar{y}, & y_{gi} > \bar{y} \\ 0, & \text{else} \end{cases},$$

the $BSA = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|\bar{y}_g - \bar{y}| = 2Cov(\bar{y}_g, I_{o\bar{y}}) = 2\sum_{g=1}^{G}\sum_{i=1}^{n_g} b_{gi}$

$$b_g = \begin{cases} \bar{y}_g - \bar{y}, & \bar{y}_g > \bar{y} \\ 0 \quad , & \text{else} \end{cases}$$

and the

$WSA = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \bar{y}_g| = \sum_{g=1}^{G} 2Cov\left(y_{gi}, I_{\bar{y}_g}\right) = 2\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{gi}$

$$w_{gi} = \begin{cases} y_{gi} - \bar{y}_g, & y_{gi} > \bar{y}_g \\ 0 \quad , & \text{else} \end{cases}$$

Let the residuals $R = t_{gi} - b_g - w_{gi}$, we have the following cases:

1. If $w = 0$, $b > 0$, $t = 0$, we have $R = -b_g = -(\bar{y}_g - \bar{y})$ then $y_{gi} \le (\bar{y}_g, \bar{y})$, $\bar{y}_g > \bar{y}$, this imply $y_{gi} \le \bar{y} < \bar{y}_g$. Similarly, if $w > 0$, $b = 0$, $t > 0$, we have $R = t_{gi} - w_{gi} = \bar{y}_g - \bar{y}$, this imply $y_{gi} > \bar{y} \ge \bar{y}_g$. These residuals must be excluded from BSA to obtain PBSA.
2. Similarly, if $w > 0$, $b = 0$, $t = 0$, we have $R = -w_{gi} = -(y_{gi} - \bar{y}_g)$ this imply $\bar{y} \ge y_{gi} > \bar{y}_g$. If $w = 0$, $b > 0$, $t > 0$, we have $R = t_{gi} - b_g = (y_{gi} - \bar{y}_g)$, this imply $\bar{y} < y_{gi} \le \bar{y}_g$. These residuals must be excluded from WSA to obtain PWSA.
3. If $w = 0$, $b = 0$, $t = 0$, we have $R = 0$ then $y_{gi} \le \bar{y}_g \le \bar{y}$. If $w > 0$, $b > 0$, $t > 0$, we have $R = 0$ then $y_{gi} > \bar{y}_g > \bar{y}$

Note that the residuals are zeros when the group mean is between the overall mean and the individuals.

Example:
Table 2 shows MAD partitions about mean for a hypothetical data. Note that $TSA = 38$, while $BSA + WSA = 36 + 18 = 54$ that overestimate TSA by $54 - 38 = 16$. On the other hand $PBSA + PWSA = 32 + 6 = 38$ that is equal to TSA.

Table 2. TSA partitions into PBSA and PWSA for a hypothetical data

g	i	x_{gi}	$TSA = \lvert y_{gi} - \bar{y}\rvert$	$BSA = \lvert \bar{y}_g - \bar{y}\rvert$	$WSA = \lvert y_{gi} - \bar{y}_g\rvert$	PBSA	PWSA
1	1	4	2	2	0	2	0
$\bar{y}_1 = 4$	2	1	5	2	3	2	3
	3	7	1	2	3	-2	3
2	1	10	4	9	5	9	-5
$\bar{y}_2 = 15$	2	20	14	9	5	9	5
3	1	2	4	4	0	4	0
$\bar{y}_3 = 2$	2	1	5	4	1	4	1
	3	3	3	4	1	4	-1
Total		48	38	36	18	32	6
$\bar{y} = 6$							

The MAD partitions about median can be shown in the same manner as MAD partitions about mean in the following theorem.

Theorem 3
The pure partitions the total sum of absolute about median into the between sum of absolute and the within sum of absolute are

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \tilde{y}| = \sum_{g=1}^{G}\sum_{i=1}^{n_g}|\tilde{y}_g - \tilde{y}|_\delta + \sum_{g=1}^{G}\sum_{i=1}^{n_g}|y_{gi} - \tilde{y}_g|_\delta \qquad (21)$$

where

$$|\tilde{y}_g - \tilde{y}|_\delta = \begin{cases} -|\tilde{y}_g - \tilde{y}| & \text{if } y_{gi} \le \tilde{y} < \tilde{y}_g \text{ or } y_{gi} > \tilde{y} \ge \tilde{y}_g \\ |\tilde{y}_g - \tilde{y}| & \text{else} \end{cases}$$

and

$$|y_{gi} - \tilde{y}_g|_\delta = \begin{cases} -|y_{gi} - \tilde{y}_g| & if\ \tilde{y}_g < y_{gi} \le \tilde{y} \text{ or } \tilde{y}_g \ge y_{gi} > \tilde{y} \\ |y_{gi} - \tilde{y}_g| & \text{else} \end{cases}$$

Note that $|\tilde{y}_g - \tilde{y}|_\delta$ is the pure between sum of absolute about median and $|y_{gi} - \tilde{y}_g|_\delta$ is the pure within sum of absolute about median, i.e.

$$TSAM = PBSAM + PWSAM$$

Example:
Table 3 shows MAD partitions about median for a hypothetical data. Note that $TSA = 89$, while $BSA + WSA = 75 + 42 = 117$ that overestimate TSA by $117 - 89 = 28$. On the other hand $PBSAM + PWSAM = 63 + 26 = 89$ that is equal to TSA.

Table 3. TSAM partitions into PBSA and PWSAM for a hypothetical data

g	i	y_{gi}	$TSAM = \lvert y_{gi} - \tilde{y} \rvert$	$BSA = \lvert \tilde{y}_g - \tilde{y} \rvert$	$WSA = \lvert y_{gi} - \tilde{y}_g \rvert$	PBSAM	PWSAM
1	1	21	0.5	1.5	1	1.5	-1
$\tilde{y}_1 = 20$	2	13	8.5	1.5	7	1.5	7
	3	22	0.5	1.5	2	-1.5	2
	4	11	10.5	1.5	9	1.5	9
	5	20	1.5	1.5	0	1.5	0
2	1	17	4.5	4.5	0	4.5	0
	2	24	2.5	4.5	7	-4.5	7
$\tilde{y}_2 = 17$	3	17	4.5	4.5	0	4.5	0
3	1	32	10.5	13.5	3	13.5	-3
$\tilde{x}_3 = 35$	2	41	19.5	13.5	6	13.5	6
	3	31	9.5	13.5	4	13.5	-4
	4	38	16.5	13.5	3	13.5	3
Total			89	75	42	63	26
$\tilde{y} = 21.5$							

4.2.1 Exact decomposition of Pietra index

The Pietra index in terms of Lorenz curve is the maximum vertical distance between the Lorenz curve, or the cumulative portion of the total income held below a certain income percentile, and the perfect equality Line, that is the 45 degree line of equal incomes. Therefore, the Pietra index is given by

$$P = \text{Max.}[F(y) - L(F(y))] = F(\mu) - L(F(\mu)) = \int_0^{\mu} y f(y)\left(1 - \frac{1}{\mu}\right) dy = \frac{\text{MAD}}{2\mu}$$

see; Lambert(1993). Therefore,

$$P = \frac{\sum_{i=1}^{n} |y_i - \bar{y}|}{2n\bar{y}} = \frac{Cov(y, I_{O\bar{y}})}{\bar{y}}$$

Using (20) the exact decomposition for the Pietra index to between and within is

$$P = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} |y_{gi} - \bar{y}|}{2n\bar{y}} = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} |\bar{y}_g - \bar{y}|_{\delta}}{2n\bar{y}} + \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} |y_{gi} - \bar{y}_g|_{\delta}}{2n\bar{y}}$$

A similar conclusion applies to MAD about median on the Lorenz curve, where

$$P_1 = \frac{\sum_{i=1}^{n} |y_i - \tilde{y}|}{2n\bar{y}} = \frac{Cov(y, I_{O\tilde{y}})}{\bar{y}}$$

is the vertical distance between the diagonal and the curve, occurring at the abscissa $F(\tilde{y}) = 0.5$; see, Pham-Gia and Hung (2001). Using (21) the exact decomposition for this measure to between and within is

$$P_1 = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} |y_{gi} - \tilde{y}|}{2n\bar{y}} = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} |\tilde{y}_g - \tilde{y}|_{\delta}}{2n\bar{y}} + \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} |y_{gi} - \tilde{y}_g|_{\delta}}{2n\bar{y}}$$

Example:

Table 4 shows an application of how this decomposition works. It can be started from wage distribution of labor force survey conducted on non-Bahraini in 2008.

Table 4 Wage distribution according to gender for non-Bahraini in 2008

Class	Range (BD)	# of Male (X_1)	# of Female(X_2)	Pietra index
1	Below 100	3760	164	P(Total)=0.42
2	100-199	765	53	between=0.07
3	200-299	267	30	within=0.35
4	300-399	146	26	
5	400-499	405	24	Using median
6	500-599	183	31	P_1(Total)=0.33
7	600-699	81	12	between=0.04
8	700-799	43	10	within=0.29
9	800-899	85	8	
10	900-999	52	12	
11	1000-1199	41	9	
12	1200-1399	31	7	
13	1400-1599	27	5	
14	1600 and over	171	231	

Source: labor source survey 2008.

The total Pietra index is 0.42 that divided to between 0.07 and within 0.35. Also the Pietra index using median is 0.33 that divided to between 0.04 and within 0.29. This shows that the most contribution to inequality back to within group wages much more than between group wages.

4.3 Beta coefficient

In finance, the beta (β) of a stock or portfolio is a number describing the correlated volatility of an asset in relation to the volatility of the benchmark that said asset is being compared to. This benchmark is generally the overall financial market and is often estimated via the use of representative indices, such as the S&P 500. The beta coefficient was born out of linear regression analysis. It is linked to a regression analysis of the returns of a portfolio (such as a stock index) in a specific period versus the returns of an individual asset in a specific year; see, Gardner, McGowan and Moeller (2010). A Beta coefficient of 1 suggests that the stock carries the same risk as the overall market and will earn market return only. A coefficient below 1 suggests a below average risk and return while a coefficient higher than 1 suggests an above average risk and return. The regression line is then called the security characteristic line (SCL)

$$R_s = \alpha + \beta R_m + \varepsilon_i$$

R_s is the monthly total return of the stock, R_m is the monthly return of the market and ε_i are independent unobservable errors obeying a zero-mean ($E(\varepsilon) = 0$), the least square method of β is

$$b_{LS} = \frac{cov(R_s, R_m)}{var(R_m)} = \frac{cov(R_s, R_m)}{cov(R_{m,}R_m)}$$

The simple LAD regression is found by choosing a pair of parameters that minimizes

$$E|\varepsilon_i| = E(w_i\varepsilon_i) = E[w_i(R_{si} - \alpha - \beta R_{mi})] \quad (22)$$

Under the condition of $E|\varepsilon_i| \geq 0$ ($E[w_i(R_{si} - \alpha - \beta R_{mi})] \geq 0$), therefore, the LAD estimate of β could be obtained when $E[w_i(R_{si} - \alpha - \beta R_{mi})] = 0$ as

$$b_{MAD} = \frac{\sum_{i=1}^{n} w_i R_{si}}{\sum_{i=1}^{n} w_i R_{mi}} = \frac{cov(R_{si}, I_{oR_{mi}})}{cov(R_{mi}, I_{oR_{mi}})} \quad (23)$$

and the weights are taken with respect to the fixed variable R_m as

$$w_i = \frac{2\,(n-1)(\hat{I}_{oi} - \bar{I})}{n^2} \text{ and } \hat{I}_{oi} = \begin{cases} 1, & R_{mi} > \tilde{R}_m \\ 0, & \text{otherwise} \end{cases}$$

Note that $\sum w = 0$. The main advantage of b_{MAD} over b_{LS} is that b_{MAD} gives less weights for the extreme ends of independent variable (R_m). To show this from (23) b can be re-written as

$$b_{MAD} = \sum_{i=1}^{n} k_i R_{si} \quad (24)$$

with LAD weights

$$k_i = \frac{w_i}{\sum_{i=1}^{n} w_i R_{mi}}$$

In least squares method the beta can be rewritten as

$$b_{LS} = \sum_{i=1}^{n} c_i R_{si}$$

with LS weights

$$c_i = \frac{R_{mi} - \bar{R}_m}{\sum_{i=1}^{n}(R_{mi} - \bar{R}_m)^2}$$

Example:

The data used in Table 5 are downloaded from the internet using Yahoo Finance website. The Table shows the return data for S&P500 and Coca-Cola for 14 months. Note that the return is computed using (close value-opening value)/opening value.

Table 5: Return of S&P500 (R_m) and Coca-Cola (R_s) for 14 months, weights using MAD and LS and beta coefficient using two methods

Months	R_m	R_s	c_i	k_i	c_iR_s	k_iR_s	
1	-1.45	1.11	-0.0011	-0.0256	-0.0012	-0.0285	
2	-0.99	-0.92	0.0015	-0.0256	-0.0014	0.0236	$b_{LS} = 0.711$
3	-8.60	-9.22	-0.0418	-0.0256	0.3855	0.2367	$b_{MAD} = 0.721$
4	1.07	-2.73	0.0132	0.0256	-0.0361	-0.0701	
5	4.75	-3.29	0.0341	0.0256	-0.1123	-0.0844	
6	-0.60	4.12	0.0037	0.0256	0.0153	-0.1058	
7	-3.48	-0.92	-0.0127	-0.0256	0.0116	0.0236	
8	-6.12	-3.86	-0.0277	-0.0256	0.1069	0.0991	
9	-0.86	-1.18	0.0022	0.0256	-0.0026	-0.0302	
10	-4.40	0.55	-0.0179	-0.0256	-0.0098	-0.0141	
11	1.48	7.46	0.0155	0.0256	0.1160	0.1915	
12	3.58	6.86	0.0275	0.0256	0.1886	0.1761	
13	1.29	3.20	0.0144	0.0256	0.0462	0.0821	
14	-3.20	-0.38	-0.0111	-0.0256	0.0042	0.0097	

Source: http://finance.yahoo.com

From Table 5 the extreme ends of R_m have less weight using MAD-comoment than using least square method.

4.4 MAD-correlation

Using the MAD-comoment representation of MAD in (7) and (8) two MAD correlation coefficients are

$$\mathrm{H}\left(Y^{(1)}, Y^{(2)}\right) = \frac{D_{\mu(v)[12]}\left(Y^{(1)}, Y^{(2)}\right)}{D_{\mu(v)[11]}\left(Y^{(1)}, Y^{(1)}\right)} \quad (25)$$

This is the ratio of the covariance between $Y^{(1)}$ and the indicator function of $Y^{(2)}$and $Y^{(1)}$ and its indicator function and

$$\mathrm{H}\left(Y^{(2)}, Y^{(1)}\right) = \frac{D_{\mu(v)[21]}\left(Y^{(2)}, Y^{(1)}\right)}{D_{\mu(v)[22]}\left(Y^{(2)}, Y^{(2)}\right)} \quad (26)$$

This is the ratio of the covariance between $Y^{(2)}$ and the indicator function of $Y^{(1)}$ and $Y^{(2)}$ and its indicator function. In general, these two measures are asymmetrical and they are necessarily not equal. Two symmetric measures of MAD correlations are defined as

$$\mathrm{H}_1 = \frac{\mathrm{H}\left(Y^{(1)}, Y^{(2)}\right) + \mathrm{H}\left(Y^{(2)}, Y^{(1)}\right)}{2} \quad (27)$$

This is the center of two asymmetric measures. Note that these measures are defined under just first moment assumption. For more details about estimation, properties of these measures and comparison with Pearson correlation coefficient; see, Habib (2011) and Elamir (2012).

5 Conclusion

By analogy with the central comoment, the MAD about mean and median is expressed as MAD-comoment in terms of covariance between a random variable and its indicator function.

There are important applications of MAD-comoment. The semi-decomposition of MAD-comoment gave an easy and simple way to obtain the coefficient of determination in LAD regression. Moreover, the exact decomposition of Pietra index of inequality is obtained by partitions total sum of absolute into the pure between sum of absolute and the pure within sum of absolute. In financial risk analysis the beta coefficient is expressed in terms of MAD-comoment. The new formula gives less weight for extreme ends than the least square coefficient. Finally, MAD-correlation coefficients are defined under just first moment assumption.

References

Bassett, J.G. and Koenker, R. (1978) Asymptotic theory of least absolute error regression. *Journal of American Statistical Association*, **73,** 618-622**.**

Bloomfield, P. and Steiger, W.S. (1983) *Least Absolute* Deviations: *Theory, Applications and Algorithms.* Birkhiiuser, Boston.

Dodge, Y. (1987) *Statistical Data Analysis Based on the L^1-Norm and Related Method.* Editor, North-Holland, Amsterdam, Editor.

Elamir, E.A.H. (2012) On uses of mean absolute deviation: decomposition, skewness and correlation coefficients. *Metron: International Journal of Statistic*, **LXX**,n.2-3, 145-164.

Gardner J.C., McGowan, C.B. and Moeller S.E (2010) Calculating the beta coefficient and required rate of return for Coca-Cola. *Journal of Business Case Studies*, **6**, 103-110.

Gorard S. (2005) Revisiting a 90-year-old debate: the advantages of the mean deviation. *British Journal of Educational Studies*, **53**, 417-430.

Habib, E.A.E (2011) Correlation coefficients based on mean absolute deviation about median. *International Journal of Statistics and Systems*, 6, 413-428.

HUBER, P. (1981) *Robust Statistics.* New York, John Wiley and Sons.

Lambert, P. (1993) *The distribution and redistribution of income: a mathematical analysis*. Manchester University Press.

Munoz-Perez J. and Sanchez-Gomez (1990) A characterization of the distribution function: The dispersion function. *Statistics and Probability Letters*, **10**, 235-239.

Pynnonen S. and Salmi T. (1994) A report on least absolute deviation regression with ordinary linear programming. *Finnish Journal of Business Economics*, **43**, 33-49.

Sefling R. and Xiao, P. (2007) A contribution to multivariate L-moments: L-comoment matrices. *Journal of Multivariate Analysis*, **98**, 1765-1781.

Tukey, J.W. (1960) A survey of sampling from contaminated distributions, In *Contributions to Probability and Statistics***,** (Edited by I. Olkin), Paper 39, pp. 448-485, The Stanford Press, Stanford, CA.