Liu Estimator Based on An M Estimator

M Tahmin Edicisine Dayalı Liu Tahmin Edicisi

Özlem ALPU,[a]
Hatice ŞAMKAR[a]

[a]Department of Statistics,
Eskisehir Osmangazi University
Faculty of Arts and Sciences,
Eskişehir

Geliş Tarihi/*Received:* 13.07.2009
Kabul Tarihi/*Accepted:* 23.10.2009

Yazışma Adresi/*Correspondence:*
Özlem ALPU
Eskisehir Osmangazi University,
Faculty of Arts and Sciences,
Department of Statistics, Eskisehir,
TÜRKİYE/TURKEY
oalpu@ogu.edu.tr

ABSTRACT Objective: In multiple linear regression analysis, multicollinearity and outliers are two main problems. In the presence of multicollinearity, biased estimation methods like ridge regression, Stein estimator, principal component regression and Liu estimator are used. On the other hand, when outliers exist in the data, the use of robust estimators reducing the effect of outliers is prefered. **Material and Methods:** In this study, to cope with this combined problem of multicollinearity and outliers, it is studied Liu estimator based on M estimator (Liu M estimator). In addition, mean square error (MSE) criterion has been used to compare Liu M estimator with Liu estimator based on ordinary least squares (OLS) estimator. **Results:** OLS, Huber M, Liu and Liu M estimates and MSEs of these estimates have been calculated for a data set which has been taken form a study of determinants of physical fitness. Liu M estimator has given the best performance in the data set. It is found as both $MSE(\hat{\alpha}_{\mathbf{LM}}) = 0.0078 < MSE(\hat{\alpha}_{\mathbf{M}}) = 0.0508$ and $MSE(\hat{\alpha}_{\mathbf{LM}}) = 0.0078 < MSE(\hat{\alpha}_{\mathbf{L}}) = 0.0085$. **Conclusion:** When there is both outliers and multicollinearity in a dataset, while using of robust estimators reduces the effect of outliers, it could not solve problem of multicollinearity. On the other hand, using of biased methods could solve the problem of multicollinearity, but there is still the effect of outliers on the estimates. In the occurence of both multicollinearity and outliers in a dataset, it has been shown that combining of the methods designed to deal with this problems is better than using them individually.

Key Words: Multicollinearity, outlier, Liu estimator, M regression

ÖZET Amaç: Çoklu doğrusal regresyon analizinde, çoklu doğrusal bağıntı ve aykırıdeğerler iki temel problemi oluşturur. Çoklu doğrusal bağıntı varlığında, ridge regresyon, Stein tahmin edicisi, temel bileşenler regresyonu ve Liu tahmin edicisi gibi yanlı tahmin metotları kullanılır. Diğer taraftan, veri setinde aykırıdeğerler var olduğunda, aykırıdeğerlerin etkisini azaltan sağlam tahmin edicilerin kullanımı tercih edilir. **Gereç ve Yöntemler:** Bu çalışmada, çoklu doğrusal bağıntı ve aykırıdeğerlere ilişkin birleşik problemle başedebilmek için M tahmin edicilerine dayalı Liu tahmin edicileri (Liu M tahmin edicileri) ele alınmıştır. Ayrıca, Liu M tahminlerini en küçük kareler (EKK) tahminlerine dayalı Liu tahminleriyle karşılaştırmak için hata kareler ortalaması (HKO) kriteri kullanılmıştır. **Bulgular:** Fiziksel yeterliliğin bileşenlerini belirlemeye yönelik bir çalışmadan alınan veri seti için EKK, Huber M, Liu ve Liu M tahminleri ve bu tahminlere ilişkin HKO değerleri hesaplanmıştır. Ele alınan veri seti için Liu M tahminleri beklentiler doğrultusunda en iyi performansı göstermiştir. Diğer bir deyişle, $HKO(\hat{\alpha}_{\mathbf{LM}}) = 0.0078 < HKO(\hat{\alpha}_{\mathbf{M}}) = 0.0508$ ve $HKO(\hat{\alpha}_{\mathbf{LM}}) = 0.0078 < HKO(\hat{\alpha}_{\mathbf{L}}) = 0.0085$ elde edilmiştir. **Sonuç:** Bir veri setinde aykırıdeğer olması durumunda sağlam tahmin edicilerinin kullanılması aykırıdeğerlerin etkisini azaltmasına karşın çoklu doğrusal bağıntı problemine çözüm getirememektedir. Diğer taraftan, yanlı tahmin tekniklerinin kullanılması da çoklu doğrusal bağıntı problemine çözüm getirebilmekte ancak aykırıdeğerlerin etkisi hala sürmektedir. Eğer veri setinde aykırıdeğer ve çoklu doğrusal bağıntı aynı anda varsa, bu problemlerle ayrı ayrı uğraşmak yerine problemler için geliştirilen yöntemleri birleştirerek kullanmanın daha faydalı olacağı bir örnek üzerinde gösterilmiştir.

Anahtar Kelimeler: Çoklu doğrusal bağıntı, aykırıdeğer, Liu tahmin edicisi, M regresyon

Turkiye Klinikleri J Biostat 2010;2(2):49-53

Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1.1)$$

where **y** is a vector of *n* response values, **X** is an *n x p* matrix of rank *p*, **β** is a vector such that $E(\boldsymbol{\varepsilon}) = 0$, and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I_n}$ .All variables in this model are corrected for their means and scaled to unit length, so that **X'X** is in correlation form.

The ordinary least squares estimator (OLS) of **β** is obtained as follows:

$$\hat{\boldsymbol{\beta}}_{\mathbf{OLS}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

If the columns of **X** are multicollinear, then the OLS estimator of **β** is an unreliable estimator due to the large variances associated with its elements.

To cope with the disadvantages of the OLS estimator in the occurrence of multicollinearity, many estimation methods have been offered. One of the most popular methods is ridge regression proposed by Hoerl and Kennard as

$$\hat{\boldsymbol{\beta}}_{\mathbf{R}} = (\mathbf{X'X} + k\mathbf{I})^{-1}\mathbf{X'X}\,\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ is the OLS estimator and *k* is the biasing parameter.[1,2]

To overcome the multicollinearity, Liu combined the Stein estimator with ridge estimator and proposed a new biased estimator as

$$\hat{\boldsymbol{\beta}}_{\mathbf{L}} = (\mathbf{X'X} + \mathbf{I})^{-1}(\mathbf{X'X} + d\mathbf{I})\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$$

where *d* is the biasing parameter.[3]

The advantage of $\hat{\boldsymbol{\beta}}_{\mathbf{L}}$ over $\hat{\boldsymbol{\beta}}_{\mathbf{R}}$ is that $\hat{\boldsymbol{\beta}}_{\mathbf{L}}$ is a linear function of *d*. So the selection of *d* is simpler than the selection of *k*. [4]

For determining biasing parameter *d*, Liu gave the estimator of *d* by minimizing the mean square error of Liu estimator as

$$\hat{d} = 1 - \hat{\sigma}^2\left[\frac{\displaystyle\sum_{i=1}^{p}\frac{1}{\lambda_i(\lambda_i+1)}}{\displaystyle\sum_{i=1}^{p}\frac{\hat{\alpha}_i^2}{(\lambda_i+1)^2}}\right]$$

where $\hat{\boldsymbol{\alpha}} = \mathbf{Q}'\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ and $\hat{\sigma}^2$ are the ordinary least squares estimates of **α** and **σ²** and **Q** is the matrix of eigenvectors corresponding to the eigenvalues of **X'X**.[3,5] In addition, many methods for appropriate *d* value in the literature have been studied by Arslan and Billor; Liu; Akdeniz and Ozturk; Ozkale and Kacıranlar.[4,6-9]

Since the Liu estimator $\hat{\boldsymbol{\beta}}_{\mathbf{L}}$ is obtained by shrinking the OLS estimator $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ using the matrix $(\mathbf{X'X} + \mathbf{I})^{-1}(\mathbf{X'X} + d\mathbf{I})$, the occurrence of outliers in the *y* direction may affect $\hat{\boldsymbol{\beta}}_{\mathbf{L}}$.[6] When outliers exist in the data, the use of robust estimators reducing the effect of outliers is prefered. The most popular of all robust estimation techniques is M estimation proposed by Huber.[10] The M estimator minimizes the following objective function

$$\min_{\beta}\sum_{\mathrm{i=1}}^{\mathrm{n}}\boldsymbol{\rho}\left(\frac{y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}}}{s}\right)$$

Differentiating the objective function with respect to the coefficients **β**, defining $\boldsymbol{\psi} = \boldsymbol{\rho}'$ and setting the partial derivates to 0, the system of equations can be written

$$\min_{\beta}\sum_{\mathrm{i=1}}^{\mathrm{n}}\psi\left(\frac{y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}}}{s}\right).\mathbf{x_i} = 0$$

where *s* is a robust estimate of scale.[11] M estimators are robust to the outliers in *y* direction.[10]

In this study, to handle with combined problem of multicollinearity and outliers, it has been examined Liu M estimator in a data set. The Liu M estimates have been compared with Liu estimates in terms of MSE.

MATERIAL AND METHODS

LIU M ESTIMATOR

The regression model given in Eq(1.1) can be rewritten in the canonical form as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where **Z=XQ, α=Q'β** and **Q** is the orthogonal matrix with columns that constitute the eigenvectors of **X'X**. Then $\mathbf{Z'Z} = \mathbf{Q'X'XQ} = \Lambda = diag(\lambda_1,\ldots,\lambda_p)$ where $\lambda_1 \geq \ldots \geq \lambda_p > 0$ are the ordered eigenvalues of **X'X**. For model in Eq(1.1), the OLS estimator $\hat{\boldsymbol{\alpha}}_{\mathbf{OLS}} = \boldsymbol{\Lambda}^{-1}\mathbf{Z'y}$ and the Liu estimator $\hat{\boldsymbol{\alpha}}_{\mathbf{L}} = (\boldsymbol{\Lambda} + \mathbf{I})^{-1}(\boldsymbol{\Lambda} + d\mathbf{I})\hat{\boldsymbol{\alpha}}_{\mathbf{OLS}}$. Note that any estimator $\hat{\boldsymbol{\alpha}}$ of **α** has a corresponding estimator $\hat{\boldsymbol{\beta}} = \mathbf{Q}'\hat{\boldsymbol{\alpha}}$

such that $MSE(\hat{\boldsymbol{\alpha}}) = MSE(\hat{\boldsymbol{\beta}})$; hence it is sufficient consider only the canonical form.

As mentioned before, since the OLS is used in Liu estimator, the presence of outliers in y direction may affect $\hat{\boldsymbol{\alpha}}_{\mathbf{L}}$. Then, the Liu M estimator of $\boldsymbol{\alpha}$ becomes

$$\hat{\boldsymbol{\alpha}}_{\mathbf{LM}} = (\boldsymbol{\Lambda} + \mathbf{I})^{-1}(\boldsymbol{\Lambda} + d\mathbf{I})\hat{\boldsymbol{\alpha}}_{\mathbf{M}}\,, \qquad (2.1)$$

where $\hat{\boldsymbol{\alpha}}_{\mathbf{M}}$ is the value obtained by the M estimator. This estimator is resistant to the combined problem of multicollinearity and outliers in the y direction.

For robust choice of d in Eq.(2.1) is given the following formula by Arslan and Billor.[6]

$$\hat{d}_M = 1 - \hat{A}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i(\lambda_i + 1)}}{\sum_{i=1}^{p} \frac{\hat{\alpha}_{Mi}^2}{(\lambda_i + 1)^2}} \right]$$

This equation can be generalized to

$$\hat{d}_{Mh} = 1 - h.\hat{A}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i(\lambda_i + 1)}}{\sum_{i=1}^{p} \frac{\hat{\alpha}_{Mi}^2}{(\lambda_i + 1)^2}} \right], \qquad h > 0$$

In practice, if d is between 0 and 1, then there is no problem is the use of $\hat{d}_M$; otherwise, it is suggested to use the generalized formula for d i.e. $\hat{d}_{Mh}$. $\hat{A}^2$ in above formulas is given as

$$\hat{A}^2 = \frac{s^2(n-p)^{-1}\sum_{i=1}^{n}\left[\psi\left(r_i / s\right)\right]^2}{\left[\frac{1}{n}\sum_{i=1}^{n}\psi'(r_i / s)\right]^2}.$$

MEAN SQUARE ERROR OF ESTIMATORS

In this study, Liu, Liu M and M estimators will be compared in terms of MSE criterion, which is the measure of the closeness of the estimate to the parameter.

The MSE of the Liu estimator is given as

$$MSE(\hat{\alpha}_{\mathbf{L}}) = \sigma^2 \sum_{i=1}^{p} \frac{(\lambda_i + d)^2}{\lambda_i(\lambda_i + 1)^2} + (d-1)^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + 1)^2}$$

where λ_i is the eigenvalues of $\mathbf{X'X}$. Liu showed that the MSE of this estimator is less than that of the OLS estimator at d values ($0<d<1$) for all values of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$.[3]

MSE for Liu M estimator is given as

$$MSE(\hat{\alpha}_{\mathbf{LM}}) = \sum_{i=1}^{p} \frac{(\lambda_i + d)^2}{(\lambda_i + 1)^2}\Omega_{ii} + (d-1)^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + 1)^2}$$

$$MSE(\hat{\alpha}_{\mathbf{M}}) = \sum_{i=1}^{p} \Omega_{ii}$$

where $\Omega = Cov(\hat{\alpha}_{\mathbf{M}})$.Arslan and Billor provided that there exists $0<d<1$ such that $MSE(\hat{\alpha}_{\mathbf{LM}}) < MSE(\hat{\alpha}_{\mathbf{M}})$ and under following conditions $MSE(\hat{\alpha}_{\mathbf{LM}}) < MSE(\hat{\alpha}_{\mathbf{L}})$:[6]

- $\boldsymbol{\psi}$ is skew-symmetric and nondecreasing,
- the errors are symmetric,
- $\Omega = Cov(\hat{\alpha}_{\mathbf{M}})$ is finite.

DATA SET

To compare the performance of the Liu M estimator with the other estimators in the presence of outliers in the y direction and multicollinearity, it has been used VO2 data set which has been taken from a study of determinants of physical fitness. The data set includes information of 233 subjects for 13 variables. In this study, totally 7 variables have been taken as a dependent variable and 6 independent variables by taking account of the cases of multicollinearity and outliers. The dependent variable is measured maximal oxygen uptake on the treadmill test (ml/kg/min), a body weight adjusted measure of the maximum amount of oxygen consumed in exercise. It is usually measured using a treadmill test protocol. The independent variables used are age (years), maximal heart rate(beat/min), duration of treadmill test(seconds), maximum systolic blood pressure on the treadmill test (mmHg), maximum diastolic blood pressure on the treadmill test (mmHg), functional aerobic impairment (percent relative to age and sex norm). The data are available on the web site (http://people.umass.edu/be640/yr2004/resources/data2002/index.html). For the analysis of the data has been used R version 2.6.2 program.

RESULTS

For this data, it is found that condition number is moderately large (106.0022). The fact that the con-

TABLE 1: OLS, Liu estimates, norms and MSE values of the estimates.

	OLS	Liu
$\hat{\alpha}_1$	-0.5456	-0.5262
$\hat{\alpha}_2$	-0.1625	-0.1674
$\hat{\alpha}_3$	0.0684	0.0712
$\hat{\alpha}_4$	0.2148	0.2087
$\hat{\alpha}_5$	0.3472	0.3353
$\hat{\alpha}_6$	0.2646	0.2597
MSE	0.0513	0.0085
$\|\hat{\alpha}\|^2$	0.5655	0.5334
	$\hat{\sigma}$ = 0.0145	*d*=0.9551

dition number is large indicates that problem of multicollinearity exists. Thus, usage of biased estimation methods instead of the OLS is preferred. In this study, to compare OLS estimates with biased estimation methods, Liu estimator from biased estimation methods is calculated and given in Table 1.

Comparison of MSE values of the OLS and Liu estimates shows that the effect of multicollinearity decreases. Moreover, a comparison of $\hat{\alpha}_{\mathbf{OLS}}$ with $\hat{\alpha}_{\mathbf{L}}$ indicates that there is a remarkable change in the estimates of the regression coefficients, since the Liu estimate produces a vector of regression coefficients with a smaller norm than does the estimate $\hat{\alpha}_{\mathbf{OLS}}$.

Whereas use of Liu estimator decreases multicollinearity problem, the effect of outliers on Liu estimates is still present if there are outliers in the data set. Thus, it is necessary to examine whether an outlier occurs in the data set. For this aim, Figure 1 has been drawn.

According to Figure 1, whereas there are several outliers (5, 31, 33, 62, 67, 168, 169, 173 and 192. observations) in *y* direction, there is no outlier in *x* direction.

When the outliers in *y* direction exist, usage of Liu M estimator will be more reliable. As mentioned before, Huber type M estimators are resistant in the presence of outliers in *y* direction. For the data set, estimates of Huber M, scale estimate and MSE of the estimates have been calculated and given below:

$$\hat{\boldsymbol{\alpha}}_{\mathbf{M}} = \begin{bmatrix} -0.5401 \\ -0.1535 \\ 0.0842 \\ 0.2016 \\ 0.3208 \\ 0.2843 \end{bmatrix}$$

$$\hat{\sigma}_{\mathbf{M}} = 0.0138$$

$$MSE(\hat{\alpha}_{\mathbf{M}}) = 0.0508$$

Due to occurrence both the multicollinearity and outliers in *y* direction in the data set, the use of Liu M estimator is preferred. For this reason, firstly calculated M regression coefficients are used to find the biasing parameter *d*. Then, Liu M estimates are obtained by using in Eq.(2.1) the calculated biasing parameter and M estimates. Liu M estimates have been given in Table 2. Moreover, Liu estimates have been also shown in the same table to compare the results.

If it is examined MSE values of Liu and Liu M in Table 2, it can be seen the slight difference in favor of Liu M. Moreover, reduction in multicollinearity of Liu estimator is not same magnitude with

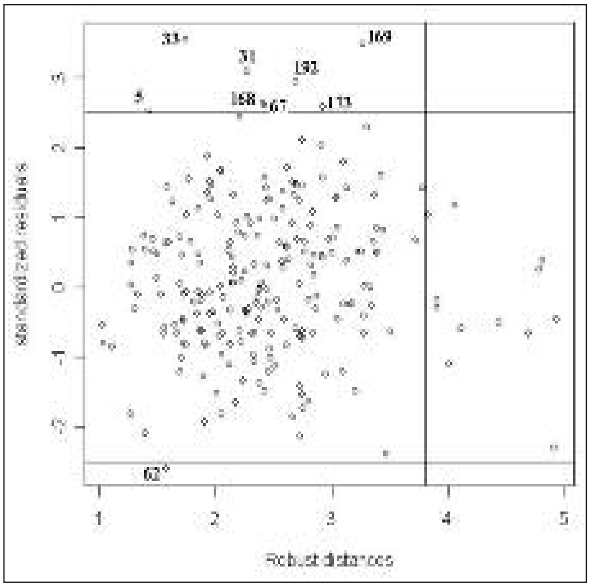

FIGURE 1: Robust residuals versus distances for VO2 data set.

TABLE 2: Liu and Liu M estimates and MSE values.

	Liu	Liu M
$\hat{\alpha}_1$	-0.5262	-0.5151
$\hat{\alpha}_2$	-0.1674	-0.1603
$\hat{\alpha}_3$	0.0712	0.0869
$\hat{\alpha}_4$	0.2087	0.1944
$\hat{\alpha}_5$	0.3353	0.3068
$\hat{\alpha}_6$	0.2597	0.2768
MSE	0.0085	0.0078
d	0.9551	0.9414

that of Liu M estimator, because the estimate of biasing parameter of Liu is closer to 1.

CONCLUSION AND DISCUSSION

From comparisons of $\hat{\alpha}_{\mathbf{OLS}}, \hat{\alpha}_{\mathbf{M}}, \hat{\alpha}_{\mathbf{L}}$ and $\hat{\alpha}_{\mathbf{LM}}$ estimates, and MSEs of these estimates the following conclusions can be drawn.

When the OLS and M estimates are compared, it is showed that the outliers in the *y* direction have a slight effect on the estimates of the parameters. The fact that the MSE of Huber M estimates is less than that of OLS is an indicator that the effect of outliers decreases. On the other hand, the multicollinearity problem can not be solved alone with the use of Huber M estimates.

Comparison of MSE values of the OLS and Liu estimates shows that the effect of multicollinearity decreases with the use of Liu estimates. But, the effect of outliers on Liu estimates is still present.

If MSE values of Liu and Liu M estimates compare with MSE values of OLS and Huber M estimates, it can be said that the effect of multicollinearity in both Liu and Liu M estimates decreases. However, Liu estimator is still affected by the outliers while the Liu M estimator improves the estimates of the parameters by reducing the effect of outliers.

When it is considered M and Liu M estimator, one can see slight differences in these estimates. There is also similar situation for OLS and Liu estimator. That's why is that *d* values used to compute Liu and Liu M estimates is close to 1. For *d*=1, $\hat{\alpha}_{\mathbf{L}}$ is equal to $\hat{\alpha}_{\mathbf{OLS}}$ and $\hat{\alpha}_{\mathbf{LM}}$ is equal to $\hat{\alpha}_{\mathbf{M}}$.

The results of MSE of all the estimators have satisfied our expectation as mentioned before. In other words, it is found as both $MSE(\hat{\alpha}_{\mathbf{LM}}) <$ $MSE(\hat{\alpha}_{\mathbf{M}})$ and $MSE(\hat{\alpha}_{\mathbf{LM}}) < MSE(\hat{\alpha}_{\mathbf{L}})$.

As a conclusion, it has been shown on a numerical example that biased estimation methods based on robust estimates should be used in the presence of outliers in *y* direction and multicollinearity in a data set. If there are outliers in *x* direction or both in *x* and *y* direction in a data set, it is suggested to use different robust methods.

REFERENCES

1. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970a;12(1):55-67.
2. Hoerl AE, Kennard RW. Ridge regression: Applications to nonorthogonal problems. Technometrics 1970b;12(1):69-82.
3. Liu K. A new class of biased estimate in linear regression. Communication in Statistics A 1993;22(2):393-402.
4. Ozkale MR, Kacıranlar S. A prediction-oriented criterion for choosing the biasing parameter in Liu estimation. Communication in Statistics- Theory and Methods 2007; 36(10):1889-903.
5. Kacıranlar S, Sakallıoglu S. Combining the Liu estimator and the principal component regression estimator. Communication in Statistics- Theory and Methods 2001;30(12): 2699-705.
6. Arslan O, Billor N. Robust Liu estimator for regression based on an M estimator. Journal of Applied Statistics 2000;27(1):39-47.
7. Liu K. Using Liu type estimator to combat collinearity. Communication in Statistics- Theory and Methods 2003;32(5):1009-20.
8. Akdeniz F, Ozturk F. The distribution of stochastic shrinkage biasing parameters of the Liu type estimator. Applied Mathematics and Computation 2005;163(1):29-38.
9. Akdeniz F, Styan GPH, Werner HJ. The general expressions for the moments of the stochastics shrinkage parameters of the Liu type estimator. Communication in Statistics- Theory and Methods 2006; 35(3):423-37.
10. Huber PJ. The basic types of estimates. In: Bradley RA, Kendall DG, Hunter JS, Watson GS, eds. Robust Statistics. 1st ed. NewYork: Wiley; 1981. p.43-198.
11. Hampel FR. Robust Statistics: The Approach Based on Influence Function. 1sted. New York: Wiley; 1986. p.105.