Model Based Macro-Editing Approach to State and Area Estimates from the Current Employment Statistics Survey[1]

Julie Gershunskaya
U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Suite 4985, Washington, DC, 20212

Abstract

Estimates of employment from the Current Employment Statistics (CES) survey are published every month for a large number of cells defined at various detailed industrial and geographic levels. Before the estimates are released for publication, they need to be reviewed. The purpose of the review is to isolate cells that may contain erroneously reported and influential records not captured during the editing procedure. The traditional approach to the macro-editing is to compare the current estimates to the historical data and mark any significant deviation as suspicious. However, it may happen that estimates deviate from the historical records for legitimate reasons (for example, due to a changing economic pattern). We propose to use a model based approach leading to a more effective screening. The model considered in this paper is the area-level Fay-Herriot model, where we apply a robust method for estimating the model parameters. A standardized difference between the sample based estimate and the synthetic part of the model predictor is used as the basis for screening. While the general CES policy is to rely on the purely sample based estimates when the sample is moderately large, a possibly useful by-product of the proposed screening procedure is the set of the robust model-based estimates that can be used to replace the direct sample estimates in a limited number of extreme cases.

Key Words: statistical data editing, macro-editing, outlier, robust small area estimation

1. Introduction

Before survey estimates produced by statistical agencies are released for publication, they need to be reviewed. Usually, the estimates are compared to analogous quantities from past years of the same survey. Large deviations from these quantities are deemed suspicious and are subjected to further analysis. A procedure for identifying unusual estimates is called the aggregation form of macro-editing. See more discussion in De Waal (2009). There are several drawbacks in the traditional aggregation method of macro screening: past quantities may deviate from the current estimates for legitimate reasons; the allowed amount of deviation from past quantities is somewhat arbitrary. The procedure may lead to biased estimates; "bending" current estimates in the direction of past years' values may result in missed "turning points."

[1] Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

In this paper, we formalize the aggregation method of macro-editing in terms of a statistical model. This allows for the opportunity to apply standard methods of model fitting and checking in exploring deviations from the assumed model.

The proposed method is based on the well-known Fay-Herriot model (Fay and Herriot 1979), which belongs to the area-level model variety employed in small area estimation (SAE). A distinguished characteristic of area-level models is that the area-level summary statistics (e.g., direct sample estimates) enter into a model as observed data. Auxiliary variables used in the model, as well, carry information at the area level. This makes area-level models suitable for application for the aggregate type of macro-editing. The traditional past years' quantities are now used as auxiliary variables in the model.

Let $\hat{Y}_i$ denote the sample based estimate in area i and let D_i be its sampling variance. Suppose a vector $\mathbf{X}_i = \left(X_{i1}, ..., X_{ip} \right)^T$ of auxiliary information is available for each area. The Fay-Herriot model is given by conditions (1) and (2) below. For each area $i = 1, ..., M$,

$$\hat{Y}_i \overset{ind}{\sim} N\left(Y_i , D_i \right), \tag{1}$$

$$Y_i \overset{ind}{\sim} N\left(\mathbf{X}_i^T \boldsymbol{\beta}, A \right), \tag{2}$$

where a p - dimensional vector of coefficients $\boldsymbol{\beta}$ and variance A are unknown parameters of the model, Y_i is the unknown true population parameter, which is the target of the estimation. Variance D_i is assumed to be known. In practice, some form of a generalized variance function is used to approximate variances.

Suppose for a moment that parameters $\boldsymbol{\beta}$ and A are known. The screening idea is simple. Form standardized values

$$r_i = \frac{\hat{Y}_i - \mathbf{X}_i^T \boldsymbol{\beta}}{\sqrt{D_i + A}}. \tag{3}$$

If the model holds, then r_i's are independent standard normal variables:

$$r_i \overset{iid}{\sim} N\left(0,1 \right) \tag{4}$$

Then, for a pre-specified threshold h, the values $\hat{Y}_i$ for which r_i exceeds the threshold in absolute value ($\left| r_i \right| > h$), are marked as outliers.

There are two general reasons for extreme values in direct estimates. It may happen that a direct estimate does not estimate the finite population parameter well; for example, it may be due to a non-sampling error or it could be caused by gross errors in micro records. We call such an estimate a "real" outlier. We might say that assumption (1) does not hold for such an area. Another possibility is that, although $\hat{Y}_i$ may be a reasonable estimate of the truth, assumption (2) does not hold for the true population parameter in area i. In such a

case, we would "falsely" mark the estimate $\hat{Y}_i$ as an outlier. Thus, a follow up analysis of the screened estimates is important.

2. Robust estimation of the model parameters

Existence of outliers points to a model failure. In other words, it contradicts the assumption that the model holds for all observations. Thus, we regard statements (1) and (2) as the *working model*, assuming it holds for the bulk of the data while allowing the possibility for extreme values in a handful of $\hat{Y}_i$'s.

Since parameters $\boldsymbol{\beta}$ and A are unknown, they have to be estimated from the data. To protect estimates of the model parameters from the effects caused by extreme values in $\hat{Y}_i$'s, we use a robust method of estimation based on bounded Huber functions.

To estimate the parameters, Fay and Herriot (1979) solve simultaneously the following equations (5) and (6):

$$\sum_{i=1}^{M} \frac{\left(\hat{Y}_i - \mathbf{X}_i^T \boldsymbol{\beta}\right)\mathbf{X}_i}{D_i + A} = 0, \tag{5}$$

$$\sum_{i=1}^{M} \frac{\left(\hat{Y}_i - \mathbf{X}_i^T \boldsymbol{\beta}\right)^2}{D_i + A} = M - p. \tag{6}$$

Instead of (5) and (6), let us solve corresponding robustified system of equations:

$$\sum_{i=1}^{M} \frac{1}{\sqrt{D_i + A}} \psi_b\left(r_i\right)\mathbf{X}_i = 0, \tag{7}$$

$$\sum_{i=1}^{M} \psi_b^2\left(r_i\right) = \left(M - p\right) E_{\Phi}\left[\psi_b^2\left(r\right)\right], \tag{8}$$

where $r_i = \dfrac{\hat{Y}_i - \mathbf{X}_i^T \boldsymbol{\beta}}{\sqrt{D_i + A}}$ and $\psi_b\left(r_i\right)$ is a bounded function; $E_{\Phi}\left[\psi_b^2\left(r\right)\right]$ is expectation of $\psi_b^2\left(r\right)$ under standard normality of random variable r. The above equations are in analogy to Huber's Proposal 2 (Huber 1964; see also Hampel *et al*. 1986, page 234.)

If $\psi_b\left(r_i\right)$ is the Huber function

$$\psi_b\left(r_i\right) = \min\left(b, \max\left(-b, r_i\right)\right), \tag{9}$$

for a predetermined fixed $0 < b < \infty$, then

$$E_{\Phi}\left[\psi_b^2\left(r\right)\right] = 2\left(\Phi\left(b\right) - 1\right)\left(1 - b^2\right) + 1 - 2b\varphi\left(b\right), \tag{10}$$

where $\Phi\left(\cdot\right)$ is the standard normal cumulative distribution function and and $\varphi\left(\cdot\right)$ is the standard normal probability density function. For example, if $b = 1.345$ then

$E_\Phi\left[\psi_b^2(r)\right]=0.7102$ and if $b=2$ then $E_\Phi\left[\psi_b^2(r)\right]=0.9205.$ In what follows, we denote $E_\Phi\left[\psi_b^2(r)\right]$ by letter c.

To solve (7) and (8), we apply the Newton-Raphson algorithm and find zeros of the following functions $f_1(\boldsymbol{\beta})$ and $f_2(A)$:

$$f_1(\boldsymbol{\beta})=\sum_{i=1}^{M}\frac{1}{\sqrt{D_i+A}}\psi_b(r_i)\mathbf{X}_i, \tag{11}$$

$$f_2(A)=\sum_{i=1}^{M}\psi_b^2(r_i)-(M-p)c. \tag{12}$$

The corresponding derivatives are

$$\frac{\partial f_1(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{M}\frac{1}{\sqrt{D_i+A}}\mathbf{X}_i^T\frac{\partial\psi_b(r_i)}{\partial\boldsymbol{\beta}}, \tag{13}$$

$$\frac{\partial f_2(A)}{\partial A}=2\sum_{i=1}^{M}\psi_b(r_i)\frac{\partial\psi_b(r_i)}{\partial A}. \tag{14}$$

The derivatives of $\psi_b(r_i)$ are

$$\frac{\partial\psi_b(r_i)}{\partial\boldsymbol{\beta}}=-\frac{1}{\sqrt{D_i+A}}\frac{\partial\psi_b(r_i)}{\partial r_i}\mathbf{X}_i, \tag{15}$$

$$\frac{\partial\psi_b(r_i)}{\partial A}=-\frac{1}{2}\frac{1}{(D_i+A)}r_i\frac{\partial\psi_b(r_i)}{\partial r_i}, \tag{16}$$

where, for the Huber function (9), we have

$$\frac{\partial\psi_b(r_i)}{\partial r_i}=I\{|r_i|\le b\}=\begin{cases}0, & |r_i|>b\\ 1 & |r_i|\le b\end{cases}. \tag{17}$$

Thus, at the *k*-th step of the Newton-Raphson algorithm, we find

$$\boldsymbol{\beta}_k=\boldsymbol{\beta}_{k-1}+\frac{\sum_{i=1}^{M}\frac{1}{\sqrt{D_i+A_{k-1}}}\psi_b(r_{i,k-1})\mathbf{X}_i}{\sum_{i=1}^{M}\frac{1}{D_i+A_{k-1}}\mathbf{X}_i^T\mathbf{X}_i I\{|r_{i,k-1}|\le b\}}, \tag{18}$$

$$A_k = A_{k-1} + \frac{\sum_{m=1}^{M} \psi_b^2\left(r_{i,k-1}\right) - (M-p)c}{\sum_{i=1}^{M} \frac{1}{\left(D_i + A_{k-1}\right)} \psi_b\left(r_{i,k-1}\right) r_{i,k-1} I\left\{\left|r_{i,k-1}\right| \le b\right\}}, \quad (19)$$

where $r_{i,k-1} = \frac{\hat{Y}_i - \mathbf{X}_i^T \boldsymbol{\beta}_{k-1}}{\sqrt{D_i + A_{k-1}}}$ and subscripts $k-1$ and k signify estimated parameter values at the respective steps of the algorithm.

3. Application to CES data

3.1 Previously used screening method and the proposed CES model

Each month, CES sample reports are used to estimate relative change in employment for the *continuous part* of the population of businesses (that is, the set of establishments that have positive employment in both previous and current months). For estimation cell i at month t, this estimate is obtained as the ratio of two survey weighted sums

$$\hat{R}_{i,t} = \frac{\sum_{j \in S_{i,t}} w_j y_{j,t}}{\sum_{j \in S_{i,t}} w_j y_{j,t-1}}, \quad (20)$$

where $y_{j,t}$ and $y_{j,t-1}$ are employment levels reported by a unit j at months t and $t-1$; $S_{j,t}$ is a set of reporting units that have positive employment in both adjacent months.

To obtain the current month estimate of the level of employment in cell i, the estimate of the relative change $\hat{R}_{i,t}$ is applied to the previous month estimated level of employment $\hat{Y}_{i,t-1}$ and a model-based net births-deaths factor, $\hat{N}_{i,t}$, is added to the result:

$$\hat{Y}_{i,t} = \hat{Y}_{i,t-1} \hat{R}_{i,t} + \hat{N}_{i,t}. \quad (21)$$

Adjustment $\hat{N}_{i,t}$ is a projected net difference of employment added by births and lost from deaths of businesses in cell i (see BLS Handbook of Methods 2004).

The recursive scheme (21) originates once a year from a known level $Y_{i,t=0}$ that is available on a lagged basis from the quarterly census of employment.

Let

$$\hat{T}_{i,t} = \frac{\hat{Y}_{i,t}}{\hat{Y}_{i,t-1}} \quad (22)$$

be the estimated relative change in employment level.

It has long been observed that monthly changes in employment have a pronounced seasonal pattern, as well as industry and geography specific character. The assumption that current sample based estimates of the monthly changes should not substantially differ

from the corresponding historical values has always been used to screen for suspicious large deviations.

The method considered in this paper is built on a similar belief. Namely, we assume that historical data correlate with the current estimates. The model based approach formalizes, refines and quantifies the old screening procedure.

Auxiliary variable $X_{i,t}$ is the relative change in employment at month *t* in cell *i*, as forecasted from the historical data. It is assumed that $X_{i,t}$ is a good predictor of the true change $T_{i,t}$. Consider a set of estimation cells $i = 1,...,M$. For example, this may be a set of States within an industrial division. (In what follows, we call the elements of this set "areas", as is customary in the SAE field.) Modeling assumptions (1) and (2) can be written as

$$\hat{T}_{i,t} \overset{ind}{\sim} N\left(T_{i,t}, D_{i,t}\right), \tag{23}$$

$$T_{i,t} \sim N\left(\beta X_{i,t}, A_t\right). \tag{24}$$

We refer to the above model as Model 1. There is a certain belief that the monthly trends have a limited tendency to change from one year to another, for the same month of a year. We consider regression through the origin. Coefficient β can be viewed as an adjustment factor that "corrects" area specific historical information (represented by $X_{i,t}$) based on the current tendency across all areas. It is expected to be close to 1 and we choose $\beta_0 = 1$ as a starting point in the Newton-Raphson algorithm.

Alternatively, we could take into account the variance of $X_{i,t}$ by using the following variation of assumption (24):

$$T_{i,t} \sim N\left(\beta X_{i,t}, a_t V_{i,t}\right), \tag{25}$$

where $V_{i,t}$ is the variance of time series prediction $X_{i,t}$ and a_t is an unknown parameter. This is Model 2.

It is important to check that the model holds for a majority of the data. For this purpose, we employ weighted normal plots considered by Dempster and Ryan (1985). We also check that the estimate of β is reasonably close to 1.

In our experience, the model usually holds. However, in the tight timeline of the CES monthly estimation, we need a backup plan for screening. In case the model fails, we think that extreme deviations from $X_{i,t}$ still need to be scrutinized. Thus, although there may be no clear linear relationship between $T_{i,t}$ and $X_{i,t}$, we suppose each individual $T_{i,t}$ to be reasonably close to $X_{i,t}$. Consider $\delta_{i,t} = \hat{T}_{i,t} - X_{i,t}$ and note that the variance of $\delta_{i,t}$ is $D_{i,t} + V_{i,t}$. Thus, we look for extreme

$$Z_{i,t} = \frac{\hat{T}_{i,t} - X_{i,t}}{\sqrt{D_{i,t} + V_{i,t}}}. \tag{26}$$

The corresponding model can be formulated as follows:

$$\hat{T}_{i,t} \overset{ind}{\sim} N\left(T_{i,t}, D_{i,t}\right), \tag{27}$$

$$X_{i,t} \overset{ind}{\sim} N\left(T_{i,t}, V_{i,t}\right), \tag{28}$$

i.e., assume $\hat{T}_{i,t}$ and $X_{i,t}$ are unbiased and mutually independent estimators of $T_{i,t}$, having variances $D_{i,t}$ and $V_{i,t}$, respectively.

We refer to (27)-(28) as Model 3. Perhaps the assumption that $X_{i,t}$ is an unbiased estimate of the truth is overly strong. On the other hand, it replaces assumptions of Model 1 or Model 2 about similarity of the areas included in the model. When there is evidence that Models 1 or 2 fail, Model 3 may be a viable alternative for screening.

3.2 Screening results examples

We now show two examples of application of Models 1-3. Consider estimates of relative changes $T_{i,t}$ in September 2011. Models were defined separately by industries, thus the subscript for industry is omitted. Index *i* represents State, *t* stands for September 2011. The parameters of Models 1 and 2 are estimated based on combined estimates from all States within industry.

Example 1. Statewide estimates of relative change, Business Services (NAICS Sectors 54, 55, 56).

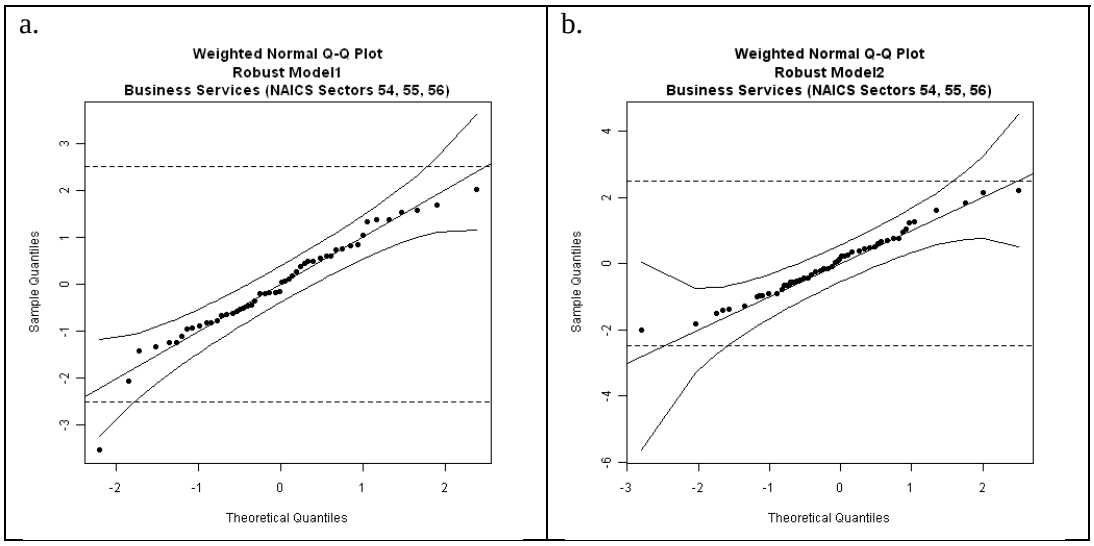

Figure 1. Weighted normal plots for Business services, for (a) Model 1 and (b) Model 2

In **Figure 1**, we show weighted normal plots for Model 1 and Model 2 based standardized residuals, along with the 95% pointwise critical regions (based on Dempster and Ryan 1985). The dashed horizontal lines are drawn at -2.5 and 2.5 levels chosen as cutoffs for "suspicious" values of the standardized residuals. These are the thresholds *h* that we alluded to in Section 1. According to the plots, both models provide reasonable fit. According to Model 1, in one State (Michigan, $i = 26$), the estimate may need further

investigation (here, $r_{26,t} = -3.5$.) Note that according to the Model 2 results, there is no values outside (-2.5,2.5) interval. This happened because the value of variance $V_{26,t}$ for $X_{26,t}$ in Michigan is relatively large, thus reducing our "trust" in $X_{26,t}$ (here, $r_{26,t} = -1.83$.)

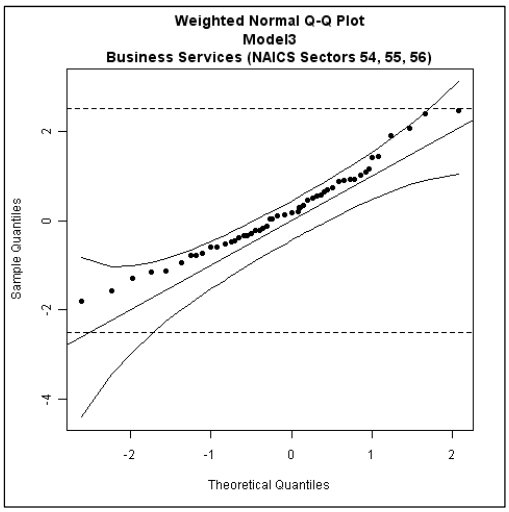

Figure 2. Weighted normal plots for Business services, Model 3

The Model 3 residuals are somewhat above the reference line (see **Figure 2**). This confirms the fact that having the adjustment factor β greater than 1 would be an improvement.

Example 2. Statewide estimates of relative change in Information (NAICS Sector 51).

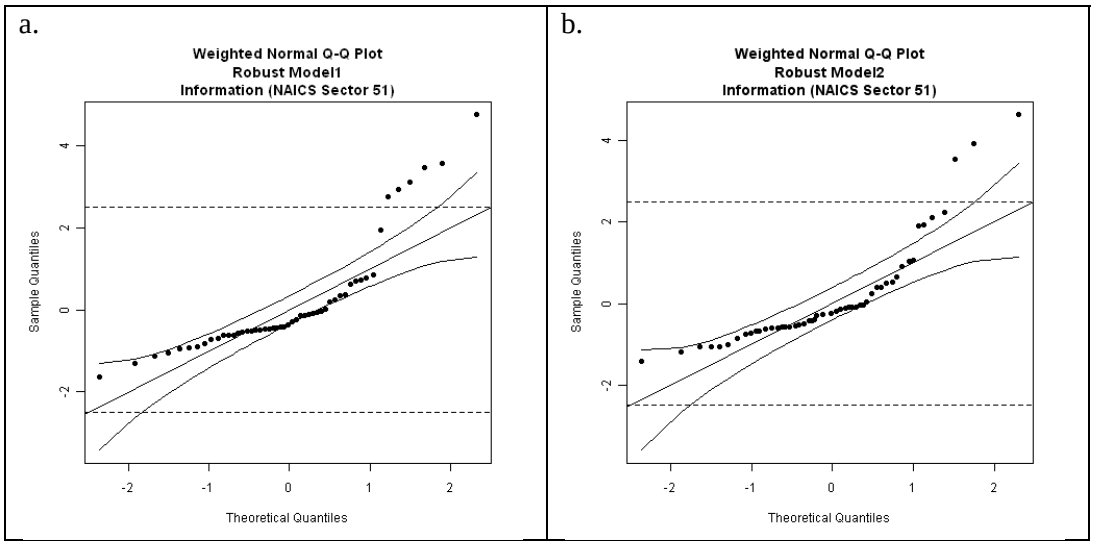

Figure 3. Weighted normal plots for Information, for (a) Model 1 and (b) Model 2

This is an example of model failure (see **Figure 3**). Apparently, forecasted values $X_{i,t}$ cannot be uniformly "adjusted" using a single factor β across all States in this industry.

The Model 3 plot (see **Figure 4**) also shows that a number of States does not fit the assumption that $\beta = 1$. There is the tendency in a group of States to be higher than corresponding values of $X_{i,t}$, rather confirming that the forecasts based on history are not good estimates of the current employment in these States and, in a sense, supporting validity of the sample based estimates.

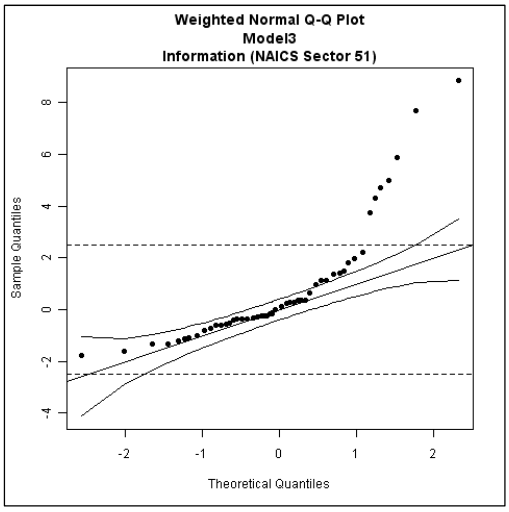

Figure 4. Weighted normal plots for Business services, Model 3

4. Summary

In this paper, we proposed a method of screening direct sample based estimates before their release for publication. The method is easy to use and it may become a convenient tool for the analysts. It is important to keep in mind that the method (as, in truth, would be any screening procedure) is based on assumptions. As with any model, the assumptions may fail. In fact, an important advantage of the proposed approach is that it makes explicit assumptions. Model checking, for example, using the normal plots, is essential. The plots also provide analysts with additional graphical tool to aid in their decisions.

References

Bureau of Labor Statistics (2004), Chapter 2, "Employment, hours, and earnings from the Establishment survey," BLS Handbook of Methods. Washington, DC: U.S. Department of Labor. http://www.bls.gov/opub/hom/pdf/homch2.pdf

Dempster, A. P. and Ryan, L. M. (1985). Weighted normal plots. Journal of the American Statistical Association, 80 845-850.

De Waal, T. (2009). Statistical Data Editing, Handbook of Statistics, Sample Surveys: Design, Methods and Applications, Eds. D. Pfeffermann and C.R. Rao, Amsterdam:Elsevier BV. Vol. 29A, Ch. 9

Fay, R.E. and Herriot, (1979). Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data, Journal of American Statistical Association, 74, 269-277

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). Robust Statistics: the Approach Based on Influence Functions. New-York, John Wiley & Sns, Inc.

Huber, P. J. (1964). Robust estimation of a location parameter, Ann. Math. Statist. 35: 73–101.