Survey Methodology, December 1997
Vol. 23, No. 2, pp. 99-107
Statistics Canada

An Adaptive Procedure for the Robust Estimation of the Rate of Change of Investment

PHILIPPE RAVALET[1]

ABSTRACT

The presence of outliers in survey data is a recurring problem in applied statistics, and the INSEE survey on industrial investment is not immune from this. The forecasting of the rate of growth of capital investment expenditures in industry therefore comes down to robust estimation of a total in a finite population. The first part of this article analyses the estimator currently used in the Investment Survey. We show that it follows a strategy of reweighting the linear estimator. But the strict dichotomy imposed between outliers – all assumed to be nonrepresentative – and other points is not fully satisfactory from either a theoretical or a practical standpoint. These flaws can be overcome by adopting a model-based approach and estimating by GM-estimators, applied to the case of a finite population. We then construct a robust adaptive procedure that determines the appropriate estimator on the basis of the residuals observed in the sample in cases where the residuals may be assumed to be symmetrical. Lastly, this method is applied to the data from the Investment Survey for the period 1990-1995.

KEY WORDS: Economic surveys; Outliers; Robust estimation; GM estimator; Adaptive procedure.

1. INTRODUCTION

Since 1952, the Institut National de la Statistique et des Études Économiques (INSEE) has been conducting an investment survey that provides estimates of the future trend of capital investment expenditures in industry, well before the National Accounts are released or the findings of exhaustive surveys are published. The estimation of the rate of investment growth is based on the declarations of some 2,500 company heads concerning their intentions to purchase capital goods.

The almost systematic presence of outliers in these data is a major problem. Outliers can seriously distort the estimate of the average growth rate and lead to unacceptable results. According to Chambers (1986), two types of outliers may be distinguished. Nonrepresentative points designate either measurement errors, which survey staff strive to correct during data collection, or unique individuals in the population. By contrast, representative outliers designate individuals which, while somewhat unusual, cannot be considered exceptional. There are undoubtedly similar individuals in the population not questioned, and the information that they contain must be integrated into the estimate.

The problem posed here is that of robust estimation of a total in a finite population with auxiliary information, a problem to which theory provides no definitive answer. Nevertheless, various techniques, reviewed in Lee (1995), can be applied. The estimation method currently used in the Investment Survey follows the logic of reweighting the linear estimator, following Hidiroglou and Srinath (1981). However, the identification and treatment of outliers are not entirely satisfactory. In particular, all outliers are assumed to be nonrepresentative, and the dichotomy between "normal" points and outliers makes the estimation quite sensitive to the choice of outliers.

The introduction of a linear superpopulation model, which describes the change in investment at the level of individuals, enables us to better assess the unusual nature of an observation and determine how representative it is. Its estimation by means of GM-estimators is then an attractive alternative to the least squares method, whose absence of bias is quite costly in terms of variance. The adjustment of the weight function depends at the outset on characteristics of the population according to criteria now well described in the literature. Since these characteristics can change not only from one stratum to another but also over time, the significance of an adaptive procedure is obvious. On the basis of a first robust estimate, we determine the appearance of the distribution of residuals, and then we choose the estimator to be used according to a predefined rule. Following Hogg, Bril, Han and Yul (1988), we construct an adaptive procedure based on indicators of tail weight and concentration estimated from the sample, since the residuals are not expected to be asymmetrical. This procedure is applied to the data from the Investment Survey for the period 1990-1995.

2. ESTIMATOR FOR THE INVESTMENT SURVEY

2.1 Estimation Principle

In a finite population $U = \{1, ..., N\}$, which here represents a stratum of the survey, a sample $s = \{1, ..., n\}$ of size n, is drawn, and $\bar{s} = \{n + 1, ..., N\}$ designates the population not questioned. Each company is questioned on

[1] Philippe Ravalet, Division des enquêtes de conjoncture, INSEE, 15 Bd. G. Péri, BP 100, 92244 MALAKOFF CEDEX.

its investment expenditures for two consecutive years $t-1$ and t, denoted respectively x and y.

Knowing the total amount X of investments for year $t-1$ in the population, we can deduce from the estimate $\hat{Y}$ of total investments for year t the average rate of change of equipment expenditures between $t-1$ and t:

$$\hat{\theta} = \frac{\hat{Y} - X}{X}.$$

To simplify the notations, we define the parameter $\Theta = 1 + \theta = Y/X$, estimated by $\hat{\Theta} = \hat{Y}/X$.

The estimator currently used in the INSEE survey draws on the ratio method, with the level of investment in $t-1$ as auxiliary information:

$$\hat{Y}_{\text{ratio}} = \frac{X}{\sum_s x_i} \sum_s y_i.$$

This estimator may be written as a weighted linear estimator:

$$\hat{Y}_{\text{ratio}} = \sum_s w_i z_i. \qquad (1)$$

In this expression, $w_i = Xx_i / \sum_s x_j$ is the weight of individual i and $z_i = y_i / x_i$ is the annual change in its investment. Such an estimator will be sensitive to the presence of outliers on both z and w. *An atypical point* will exhibit a change z that is very different from that of the others, while an *influential point* will have a weight w that is large enough to attract, by leverage, the average rate of change of the stratum towards its own rate of change. Since the decisive criterion for characterizing an observation as an outlier is that the product wz is large enough to distort the estimate $\hat{Y}_{\text{ratio}}$, the distinction between atypical points and influential points is, of course, arbitrary. The generic term *large investors* (or LI for short) will designate these outliers as a group, while the term *extrapolatables* will refer to the other individuals in the sample.

Having carried out an *a posteriori* partition of the sample $s = \{\text{LI}\} \cup \{\textit{extrapolatables}\}$, we estimate the total investments of the rest of the population $\bar{s}$ on the basis of the behaviour of only the extrapolatable individuals according to the ratio method:

$$\hat{Y}_{\text{LI}} = \sum_s y_i + \left(\sum_{\bar{s}} x_i \right) \frac{\sum_{\{\text{extra}\}} y_i}{\sum_{\{\text{extra}\}} x_i}. \qquad (2)$$

In (2), the weight of the extrapolatables $1 + \sum_{\bar{s}} x_i / \sum_{\{\text{extra}\}} x_i$ is quite strictly greater than the weight of the large investors, which is equal to 1.

2.2 Selection of Large Investors

The large investors are selected within each stratum on the basis of their influence on the estimation of Θ according to an iterative procedure. At the outset, all individuals are assumed to be extrapolatable, and for each of them we calculate a not-taken-into-account index, measuring the impact on $\hat{\Theta}$ of its exclusion from the sample, $\text{NTIA} = (\hat{Y}^i_{\text{LI}} - \hat{Y}_{\text{LI}})/X$ where $\hat{Y}^i_{\text{LI}}$ is the estimated total without individual i.

The firm with the largest NTIA index in absolute value is said to be a large investor. $\hat{Y}_{\text{LI}}$ is then re-estimated with this new partition of U, and then the next large investor is identified. The selection stops when all extrapolatable individuals' have an influence on the estimate that is below a given threshold. The greater the number and mass of observations, the easier it is to verify this condition. Conversely, it will prove impossible to verify the condition if the number of individuals is too small; in that case, the survey manager merely makes sure that no individual has a much greater influence than the others, thus introducing an element of subjectivity into the procedure.

By this iterative mechanism, the usual phases of detection and treatment of outliers are carried out simultaneously. The main problem is that the status of an individual is not an intrinsic characteristic but instead depends on the composition of the sample. This can change from one survey to another. In addition, in certain hypothetical cases (Ravalet 1996), this procedure can lead to the unnecessary exclusion of some individuals, since at no point is the status of large investor called into question.

2.3 Strategy for Reweighting the Linear Estimator

The estimator LI in fact follows from the strategy for reweighting the linear estimator (1) presented by Hidiroglou and Srinath (1981) using the example of estimation of a total without auxiliary information. Having already carried out a partition $s = s_1 \cup s_2$ of the sample distinguishing the outliers s_1 (numbering n_1) from the other observations s_2, the authors propose to reduce, in $\hat{Y} = (N/n)\sum_s y_i$, the weight N/n of the outliers to a lower value λ by positing

$$\hat{Y}_\lambda = \lambda \sum_{s_1} y_i + \frac{N - \lambda n_1}{n - n_1} \sum_{s_2} y_i$$

and

$$\hat{Y}_\lambda = \sum_s y_i + \frac{N-n}{n-n_1} \sum_{s_2} y_i + n_1(\lambda - 1)\left[\frac{1}{n_1} \sum_{s_1} y_i - \frac{1}{n - n_1} \sum_{s_2} y_i \right].$$

The optimal value of λ that minimizes the mean square deviation of this estimator, whether or not conditional on the number of outliers in the sample, depends on several parameters of the population. Without prior information, the choice of λ is a delicate one.

Applied to the case of the estimator of the ratio with auxiliary variable x, this is written as:

$$\hat{Y}_{\text{ratio}\,\lambda} = \sum_s y_i + \sum_{\bar{s}} x_i \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} +$$

$$(\lambda - 1)\left(\frac{\sum_{s_1} y_i}{\sum_{s_1} x_i} - \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i}\right)\sum_{s_1} x_i. \quad (3)$$

The first two terms of the second member of (3) form an estimate of the total Y, under the implicit hypothesis that all outliers are in the sample, and the third is a correction taking account of the possible presence of outliers in the population not questioned. This correction is a function of the λ selected and the difference in average behaviour between the two types of individuals estimated in the sample.

When (2) and (3) are considered together, it may be seen that the estimator LI is formally equivalent to the case $\lambda = 1$. The use of $\hat{Y}_{\text{LI}}$ thus implicitly assumes that the outliers have been correctly identified and are all non-representative. In Ravalet (1996), it was shown that these two hypotheses were unfortunately seldom verified in the context of the Investment Survey.

Since the identification procedure is manual and the criterion used is relatively *ad hoc* in the absence of any hypothesis on the population, it is not impossible that some outliers will escape selection. The use of the ratio on the extrapolatables then poses the problem of the robustness of the estimation in relation to the choice of large investors. In addition, it is unlikely that all these points are unique. The atypical points, which are especially numerous among small and medium-sized firms, should instead be considered as representative. However, choosing $\lambda > 1$ would inevitably raise the question of the robustness of the third term of (3).

To try to compensate for these defects, changes to the estimator $\hat{Y}_{\text{LI}}$ are possible. For example, the mean of the extrapolatables may be replaced by a more robust estimator, and only the nonrepresentative points are designated as large investors. This technique fits into the more general framework of M-estimators, in which the existence of a model facilitates both the detection and treatment of outliers (Lee 1995). It is then no longer a matter of constructing a strict dichotomy between outliers and other points but rather of defining areas of varying representativeness.

3. ROBUST ESTIMATION BY GM-ESTIMATORS

3.1 The Linear Model and GM-Estimators

Assume the existence of a linear model ξ that links together, for the overall population U, investments x and y on dates $t-1$ and t.

$$\xi: y_i = \beta x_i + \epsilon_i$$

with

$$E(\epsilon_i) = 0$$
$$E(\epsilon_i \epsilon_j) = 0 \qquad \forall\, i \neq j.$$
$$V(\epsilon_i) = \sigma^2 \eta(x_i)$$

Slope β of the regression line passing through the origin in the superpopulation model is interpreted as the rate of change Θ in the population. The variance of y is assumed to be an increasing function of x and η is generally a power function: $\eta(x_i) = x_i^{\gamma}$.

According to the model, the best unbiased linear estimator (Brewer 1963 and Royall 1970) of the total is $\hat{Y}_{mc} = \sum_s y_i + \hat{\beta}_{mc} \sum_{\bar{s}} x_i$ where $\hat{\beta}_{mc} = (\sum_s x_i y_i / \eta(x_i))/(\sum_s x_i^2/\eta(x_i))^{-1}$ is the least squares estimator.

In the particular case $\eta(x) = x$, this expression reduces to $\hat{\beta}_{mc} = \sum_s y_i / \sum_s x_i$, estimator of the ratio. This unbiased estimator is effective only under the hypothesis of normality of the residuals, and it does not prove to be very robust.

The M-estimators (Huber 1981) serve to define a robust version of the least squares by replacing the square function, in the minimization program, with a function ρ that increases less rapidly:

$$\text{Min} \sum_s \rho\left(\frac{y_i - \beta_R x_i}{\sigma\sqrt{\eta(x_i)}}\right).$$

The M-estimator $\hat{\beta}_R$ is the solution of the following implicit equation:

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\sigma\sqrt{\eta(x_i)}}\right)\frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

where

$$\psi(t) = \frac{\partial \rho(t)}{\partial t}.$$

The function ψ, like Huber's function $\psi(t) = \text{Max}(-c, \text{Min}(t, c))$, depends on one or more adjustment constants c controlling the portion of observations that must be considered as outliers. This estimator will still be sensitive to the effect of outliers on the explanatory variable x. Therefore a more general class of estimators, called GM-estimators (Hampel, Ronchetti, Rousseeuw and Stahel 1986), is defined by means of the following implicit equation:

$$\sum_s w\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right)\psi\left(\frac{r_i}{\sigma} v\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right)\right)\frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

with

$$r_i = \frac{y_i - \hat{\beta}_R x_i}{\sqrt{\eta(x_i)}}.$$

A choice usually made is Mallows' formulation: $v(t) = 1$ and $w(t) = 1/t$. Hence a robust estimator $\hat{\beta}_R$ will verify the implicit equation

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\sigma\sqrt{\eta(x_i)}}\right) = 0. \quad (4)$$

In general, the parameter σ is unknown and must be replaced in this expression by a robust estimate $\hat{\sigma}$ of the dispersion of the residuals

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = \sum_i \psi\left(\frac{r_i}{\hat{\sigma}}\right) = 0.$$

The estimator of the total will then be:

$$\hat{Y}_{\beta R} = \sum_s y_i + \hat{\beta}_R \sum_{\bar{s}} x_i. \quad (5)$$

This estimator is studied by Gwet and Rivest (1992). In general, it is not unbiased in relation to the sample design. Chambers (1986) proposes to correct that bias by introducing into (5) a third term that estimates it robustly:

$$\hat{Y}_{\text{Chambers}} = \sum_{i\in s} y_i + \hat{\beta}_R \sum_{i\in\bar{s}} x_i +$$

$$\left(\sum_{i\in s} \frac{x_i/\hat{\sigma}\sqrt{\eta(x_i)}}{\sum_{j\in s} x_j^2/\hat{\sigma}^2\eta(x_j)} \psi_E\left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right)\right) \sum_{i\in\bar{s}} x_i.$$

Choosing a bounded function ψ_E seems a good compromise between estimator bias and variance For example, Welsh and Ronchetti (1994) opt for a Huber's function with a large adjustment constant $c = 15$. But the adjustment of ψ_E, without prior information on the density of the outliers, is always difficult.

3.2 Choice of Estimator

The desirable properties of ψ functions are now well known with reference to the problem of estimating a central tendency. They must be bounded, continuous, and equivalent to an identity in the vicinity of zero. Strictly monotone functions (Huber) are distinguished from redescending functions such as Tukey's biquadratic function, Andrew's sine and the Hampel or Cauchy function. Because their influence function tends toward zero, these estimators will be less sensitive to the presence of outliers than the Huber function. The speed of convergence toward zero is an essential characteristic of redescending functions. Those that are nil at a finite distance (Hampel, Tukey or Andrew) exclude outliers from the estimation of β, whereas the others assign them low representativeness.

The choice and adjustment of the ψ function are difficult. They greatly depend on the nature of the data and more specifically on the distribution of the residuals (Hoaglin, Mosteller and Tukey 1983, Ch. 11). An idea, however approximate, of the appearance of the distribution of the residuals should make it possible to better target both the choice and the adjustment of the estimator, and hence to make the estimation more efficient. This intuitive remark is at the origin of adaptive procedures, presented in particular by Hogg (1974) and (1982). The idea is to evaluate the nature of the distribution of the residuals, calculated on the basis of an initial robust estimate (of the norm L_1 type, for example), using carefully selected robust indicators (tail weight, asymmetry, concentration, *etc.*). The existence of these indicators makes it possible, using a predefined decision rule, to select the appropriate estimator for this situation, and the implicit equation (4) is solved by taking the first robust estimate of β as an initial value.

The idea of an adaptive procedure appears all the more attractive since it systematizes the study that must precede the choice and adjustment of an estimator. That study can prove extremely costly if it must be performed manually for each stratum of the sample and repeated for each survey.

4. CONSTRUCTION OF AN ADAPTIVE PROCEDURE

This section describes the construction of an adaptive procedure for calculating the average rate of change of investment on the basis of economic survey data. Consequently, certain choices were made in light of the specific nature and characteristics of those data and are not necessarily transposable to other regression models. In particular, after checking the data, we adopted the hypothesis of a symmetrical distribution of residuals and we excluded the case of light-tail distributions.

The construction of an adaptive procedure, which draws on the works of Moberg, Ramberg and Randles (1980), is carried out in several stages. The first step is to choose the ψ function (or family of functions) to be used. The second is to select the various criteria for characterizing the distribution of residuals. Using these criteria, a classification rule is constructed. Finally, each class is matched with the adjustment of the estimator to be used.

4.1 Choice of ψ Function

Since Huber-type monotone functions do not provide sufficient protection against outliers, only redescending functions were considered. Among them, we selected the generalized Cauchy function (used in particular by Moberg *et al.* 1980 to approximate generalized lambda functions) and the Tukey biquadratic function:

$$\psi_c(r) = \frac{cr}{(b + r)^2 + c}, \quad \forall r$$

and

$$\psi_T(r) = \frac{r}{c}\left(1 - \frac{r^2}{c^2}\right)^2, \quad \forall |r| \le c.$$

These two estimators are quite different in their treatment of outliers (see Figure 1). The biquadratic function equals zero for longer than the Cauchy function, but on the other hand it has a finite rejection point: the residuals beyond $c^*\sigma$ do not enter into the estimate, whereas the Cauchy function assigns them a certain representativeness. The parameter b serves, in principle, to control the asymmetry of ψ according to that of the residuals.

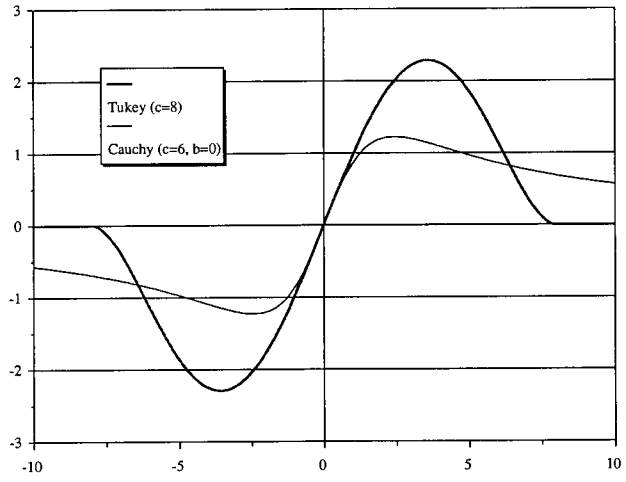

Figure 1. Cauchy and Tukey Functions

4.2 Parameter of Scale, Calculation Algorithm and Selection Criteria

In general an estimator $\hat{\sigma}$ of dispersion is defined by an implicit equation $\sum\chi(r_i/\hat{\sigma}) = 0$, where χ is an even function. It is therefore a matter of solving the system of non-linear equations in $(\hat{\beta}, \hat{\sigma})$ following:

$$\begin{cases} \sum_i \psi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0 \\ \sum_i \chi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0. \end{cases} \tag{6}$$

Rivest (1989) offers several examples showing that resolving system (6) can pose problems, owing to the fact that there may be a number of solutions, even in the case of a monotone ψ function. Following his recommendations, we will proceed in two stages. First, the parameter of dispersion σ is estimated using the median of the absolute values (MAD) of the residuals defined on the basis of the median of the individual rates of change. Then β is calculated by (4) using the value of σ found previously.

For solving (4), we preferred the reweighting algorithm to the Newton-Raphson algorithm, since it seems to converge more easily, especially when the adjustment constant is small.

Since the effectiveness of an adaptive procedure depends on the effectiveness of the decision-making process, the greatest attention must be paid to the nature, quality and robustness of the information that guides the choice of the estimator. Tail weight is an indispensable indicator, since it provides information on the relative significance of outliers in the sample and thus in the population (see Hoaglin *et al.* 1983, ch. 10). For the tail weight indicator, we adopted the proposal of Hogg (1974):

$$\tau(p) = \frac{\bar{U}(p) - \bar{L}(p)}{\bar{U}(0.5) - \bar{L}(0.5)}$$

$\bar{U}(p)$ (resp. $\bar{L}(p)$) is the mean of the np largest (resp. smallest) order statistics, using a linear interpolation when np is not whole. We chose $p = 0.05$; for the normal distribution $\tau(.05)$ is equal to 2.59.

In addition, like Hogg *et al.* (1988), we considered it important to test for the possible presence of a distribution of the double exponential type, measuring the concentration of residuals by the following *pk* indicator:

$$pk = \frac{\bar{X}(1 - \beta, 1 - \alpha) - \bar{X}(\alpha, \beta)}{\bar{X}(.5, 1 - \beta) - \bar{X}(\beta, .5)}$$

where $\bar{X}(a, b)$ is the means of the order statistics between the *na*-th and the *nb*-th, with the sizes interpolated if *na* or *nb* are not integers. We selected $\alpha = 0.05$ and $\beta = 0.15$, or $pk = 2.7$ for a normal distribution.

Finally, different studies (Moberg *et al.* 1980, Hogg *et al.* 1988) have emphasized the importance of the dissymmetry of distributions. When there are asymmetrical residuals, the bias of robust estimators can be sizable, making it tricky to use them (Chambers et Kokic 1993). In the INSEE Investment Survey, the residuals are theoretically asymmetrical since they are confined to a limited range ($r = y - \beta x \ge -\beta x$). However, we noted empirically that this asymmetry was very slight and could safely be ignored. The failure of the correction of a possible bias by the function ψ_E in Chambers' estimator moreover confirms this observation. Only the symmetrical case is considered here; the bias of the estimators defined by (5) is therefore nil.

4.3 Classification of Distributions and Adjustment of the Estimator

The definition of the decision rule was based on the study of eight specific symmetrical distributions illustrating various tail weight and concentration situations (see Table 1). We were interested in the family of contaminated distributions $CN(\alpha, K)$, with the distribution function $F(x) = (1 - \alpha)\Phi(x) + \alpha\Phi(x/K)$ where Φ is the cumulative function of the distribution $N(0, 1)$, since these distributions give a good representation of real data (Hoaglin *et al.* 1983, ch. 10), especially the data in the Investment Survey (Ravalet 1996). While Gaussian in the middle, they nevertheless contain more outliers than the normal distribution $N(0, 1)$.

Table 1
Eight Specific Distributions

		$\tau(.05)$	pk
1	Normal distribution	2.59	2.76
2	Contaminated dist $CN(.05, 3)$	2.94	2.83
3	Double exponential dist.	3.28	3.41
4	Contaminated dist $CN(.05, 10)$	4.47	2.85
5	Contaminated dist $CN(.10, 10)$	5.42	3.05
6	Contaminated dist $CN(.20, 10)$	5.64	4.44
7	Slash distribution	7.65	4.19
8	Cauchy distribution	7.82	4.78

The two indicators $\tau(0.5)$ and pk were simulated over these eight distributions, for several sample sizes. The graph of $(\tau(0.5), pk)$ serves to distinguish four groups of distributions: light-tailed, relatively unconcentrated distributions of the normal type or $CN(.05,3)$; heavy-tailed distributions of the type $CN(.05,10)$, $CN(.10,10)$, and $CN(.20,10)$, and very heavy-tailed distributions of the Slash or Cauchy type; and concentrated distributions such as the double exponential distribution. These four classes are defined (see Figure 2) by the following equation boundaries:

$$\text{Class I:}\quad \tau(0.5) \le 3.6 - \frac{14}{n} \quad \text{and} \quad pk \le 3.20$$

$$\text{Class II:}\quad 3.6 - \frac{14}{n} < \tau(0.5) \le 5.8 - \frac{35}{n}$$

$$\text{Class III:}\quad 5.8 - \frac{35}{n} < \tau(0.5)$$

$$\text{Class IV:}\quad \tau(0.5) \le 3.6 - \frac{14}{n} \quad \text{and} \quad pk > 3.20$$

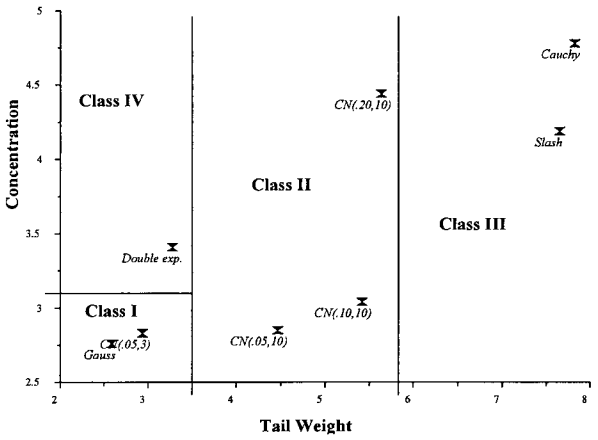

Figure 2. Four Classes of Distributions

The final stage consists in setting the adjustment of the two estimators in each class. Since we are interested only in the symmetrical case, the b parameter of the Cauchy function is nil. By simulations, we determined for the eight reference distributions the optimal constants c of the Tukey and Cauchy functions (*i.e.*, minimizing the variance of these estimators or, what amounts to the same thing here, their mean square deviation). These do indeed diminish with tail weight, except of course for the case of the double exponential distribution, which requires an adjustment similar to those used for the Slash and Cauchy distributions.

Tukey's estimator is more efficient on the normal or contaminated distributions, but it generally requires finer adjustment. Figure 3 shows the example of the contaminated distribution *CN(.10,10)*. Lastly, while the choice of the constant appears to be relatively critical for the heavy-tailed or concentrated distributions, a wide band of value is possible for distributions close to the normal distribution.

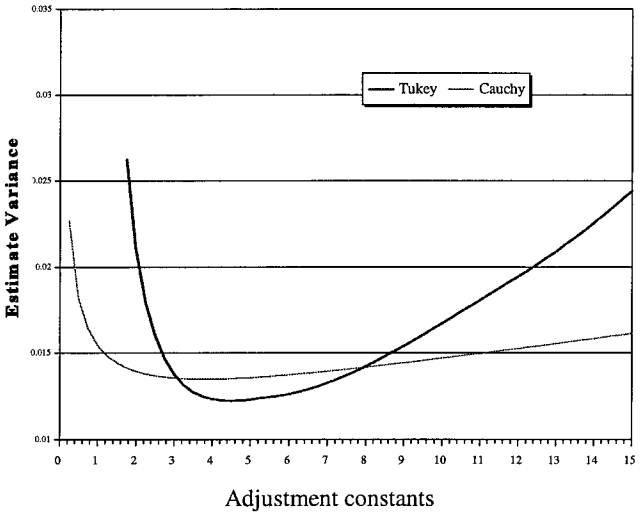

Figure 3. Variance of Tukey and Cauchy Estimators for the Distribution CN(.10,10) (n=100)

The synthesis of these results serves to define the adjustments to be used on each distribution class. These adjustments, established for samples of size 100 (Table 2), remain entirely acceptable for samples sizes between 50 and 150.

Table 2
Adjustment of Estimators by Class of Distribution of Residuals (n = 100)

Class	Tukey	Cauchy
I	7	7
II	4.5	4
III	3	1
IV	3	1

5. APPLICATION TO THE INVESTMENT SURVEY

5.1 The Problem of Stratification

The strata used for the LI estimator are defined by the cross-tabulation of an activity (18 manufacturing sectors) and a company size class (small, medium or large). Among these 54 strata, approximately 20 never contain more than 20 observations. This stratification is therefore too fine for the adaptive procedure to be used correctly, as it assumes a minimum number of observations.

Since small firms are fairly distinct from medium-sized and large firms in terms of dispersion and residuals tail weight, differentiation by size is maintained. Sectors must thus be grouped. We decided not to adopt the method used by Sohre (1995), which consists of grouping after data collection those sectors having the closest parameters (here the average change in investment). Proximity is impossible to assess in small strata, and the groups obtained are likely to change from one survey to another, making comparisons difficult. We preferred to redefine 15 new strata based on a higher classification level distinguishing only four sectors: intermediate goods, professional capital goods, automobile, and consumer goods.

5.2 Characteristics of Strata

The hypothesis of a variance of residuals independent of x in the model ξ cannot be accepted. The choice of γ in the function η is made in such a way that the curve of the residuals (in absolute value) as a function of the regressor, smoothed by the LOESS method, shows no trend (Cleveland 1979). For the stratum representing intermediate goods and medium-sized companies in the April 1995 survey (see Figure 4), $\gamma = 1.3$ is an acceptable compromise between the appearance of a downward trend for small values of x and the cancellation of the upward trend for the larger values of x. A similar examination on the other strata confirmed this choice for the manufacturing industry as a whole.

In each stratum, the distribution of the residuals systematically exhibits a heavier tail than the normal distribution, without being extremely heavy-tailed. Within a given sector, the tail weight indicator decreases with company size. The great majority of the strata representing small and medium-sized firms were assigned to Class 2. Large firms more often exhibit somewhat heavy-tailed distributions, close either to the normal distribution (Class 1), or the double exponential distribution (Class 4). Class 2 is by far the largest and represents 75% of cases. Only 20% of the distributions are recognized as somewhat heavy-tailed and are assigned in equal proportions to classes 1 and 4. On the other hand, very heavy-tailed distributions (Class 3) are unusual (less than 5% of the cases). While there appears to be a certain persistence to the classification, it is not perfect. And the changes are quite real, since they resist a slight modification of the boundaries between classes. Thus this perfectly justifies the use of an adaptive procedure.

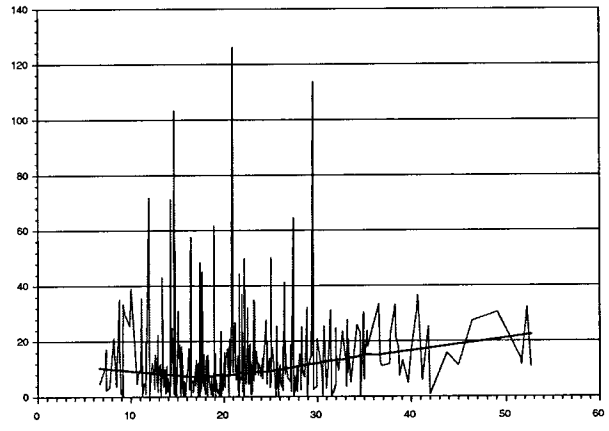

Figure 4. Absolute Value of Residuals ($\gamma = 1.3$, Intermediate Goods, Size 2, April 95)

5.3 Resulting Estimates

The estimation procedure based on (5), applied to the six surveys covering the period 1990-1995, yielded the results shown in Figure 5. Also shown are National Accounts estimates, those obtained with the LI estimator, and those from the Annual Business Survey (ABS), which is exhaustive.

For the manufacturing sector as a whole, the results of the adaptive procedure are comparable to those obtained with the LI estimator. The biquadratic function results in estimates that are consistently lower than those obtained with the Cauchy function. With a finite rejection point, the Tukey function is less influenced by the slight asymmetry toward the right in the distribution of the residuals. These new estimates are closer to those of the ABS than to the National Accounts estimates. This is hardly surprising, considering the excellent correlation between individual

ABS data and the responses obtained in the survey. As yet there is no explanation for the differences in 1991 and 1994 in relation to the National Accounts estimates. Apart from the year 1994, the estimates obtained with the Cauchy function are entirely acceptable in the intermediate goods and automobile sectors and to a lesser extent in the professional capital goods sector. On the other hand, in consumer goods, the results are fairly far from the National Accounts estimates. Here we are likely running up against a problem of sample quality. This sector is quite heterogeneous, and a few activities such as printing are poorly covered by the survey.

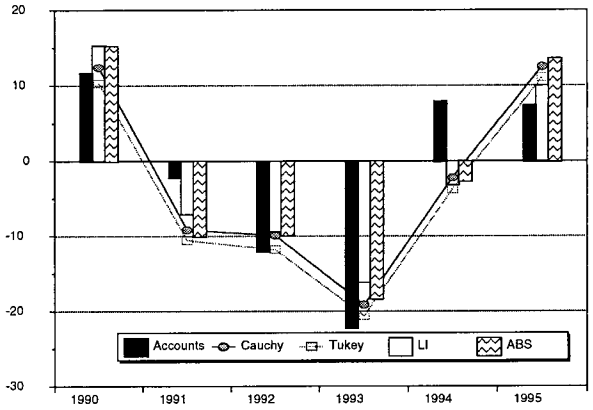

Figure 5. Investment Growth Rate in Value in the Manufacturing Industry

6. CONCLUSIONS

This article presents a theoretical justification of a procedure currently used to process data from the Investment Survey; in particular it offers a justification of the principle of excluding outliers or large investors. However, the strategy of reweighting the linear estimator following Hidiroglou and Srinath (1981) shows itself to be insufficient for this purpose in several respects, mainly having to do with the identification and treatment of representative outliers. The dichotomy between extrapolatable individuals and large investors appears too radical and leads to a lack of robustness, since the influence curve of this estimator is not continuous.

On the other hand, the hypothesis of a linear superpopulation model and its estimation by GM-estimators seemed to us to be of great interest from both a methodological and practical standpoint. The insertion of these techniques into an adaptive procedure also makes it possible to have a robust estimator for a variety of situations. Following principles described in the literature, the procedure proposed here uses indicators of tail weight and concentration of the residuals in the linear model calculated from the sample, to decide on the adjustment of the weight function to be used, it being assumed that the residuals are symmetrical. The estimates made with the Cauchy function yielded satisfactory results on the manufacturing industry, and they largely validate previously published results. The advantages of this method over the one currently used basically have to do with lower implementation costs and greater control over the methodology employed.

The adaptive procedure was constructed independently of the survey, and therefore there is no guarantee that the classification is optimal for the strata content. Furthermore, we did not study the robustness of the rule for assigning values to a class. This issue is important when one carries out several successive measurements and one wants to interpret the revisions. Clearly, further research on these classification methods is required, in order to integrate additional information such as the information yielded by earlier estimates or comprehensive surveys of the population studied.

ACKNOWLEDGEMENTS

The author wishes to thank Michel Hidiroglou and Dominique Ladiray for their comments and suggestions during the preparation of this article.

REFERENCES

BREWER, K.R. (1963). Ratio estimation and finite population: some results deducible from the assumption of an underlying stochastic process. *The Australian Journal of Statistics*, 5, 93-105.

CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

CHAMBERS, R.L., and KOKIC, P.N. (1993). Outlier robust sample survey inference. *Bulletin of the International Statistical Institute, Proceedings of the 49th Session, Book 2*, 55-72.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

GWET, J.P., and RIVEST, L.P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

HAMPEL, F.R., RONCHETTI, E., ROUSSEEUW, P.J., and STAHEL, W.E. (1986). *Robust Statistics: The Approach Based on Influence Function*. New York: John Wiley.

HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

HOAGLIN, D.C., MOSTELLER, F., and TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

HOGG, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of American Statistical Association*, 69, 909-923.

HOGG, R.V. (1982). On adaptive statistical inferences. *Communication in Statistics*, 11, 2531-2542.

HOGG, R.V., BRIL, G.K., HAN, S.M., and YUL, L. (1988). An argument for adaptive robust estimation. *Probability and Statistics Essays in Honor of Franklin A. Graybill.* Amsterdam: North-Holland/Elsevier, 135-148.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.

LEE, H. (1995). Outliers in business surveys. In *Business Survey Methods*. New York: John Wiley.

MOBERG, T.F., RAMBERG, J.S., and RANDLES, R.H. (1980). An adaptive multiple regression procedure based on M-estimators. *Technometrics*, 22, 213-224.

RAVALET, P. (1996). L'estimation du taux d'évolution de l'investissement dans l'enquête de conjoncture: analyse et voie d'amélioration. Document de travail de l'INSEE Méthodologie Statistique, 9604.

RIVEST, L.P. (1989). De l'unicité des estimateurs robustes en régression lorsque le paramètre d'échelle et le paramètre de régression sont estimés simultanément. *Canadian Journal of Statistics*, 17, 141-153.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SOHRE, P. (1995). The Adaptive KOF Procedure for the Estimation of Industry Investment. 22nd CIRET Conference, Singapore.

WELSH, A.H., and RONCHETTI, E. (1994). Bias-Calibrated Estimations of Totals and Quantiles From Sample Surveys Containing Outliers. Technical Report, Dept. of Econometrics, University of Geneva, Switzerland.