

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (v): Editing based on results (post-editing)

LOCAL AND GLOBAL SCORE FUNCTIONS IN SELECTIVE EDITING

Invited Paper

Prepared by Dan Hedlin, Statistics Sweden

I. INTRODUCTION

1. With selective editing the incoming questionnaires are prioritised, and those that have been given priority are selected for editing. The prioritisation step often involves the computation of an item score (local score) for each data value that reflects the importance of investigating this data value. A score may be computed for each item on the questionnaire, and the item scores may be combined to a unit score (global score). The paper discusses ways of summarising item scores to a unit score making use of the mathematical concept of distance.

2. A number of successful applications of selective editing have been presented at previous UNECE editing workshops. Various forms of score functions have proven useful in selective editing. Often the local score is designed so as to reflect the difference between the estimate you obtain using the reported value and the estimate you would have got with the true value. Most National Statistical Institutes want to check and edit the whole form rather than each item separately. To this end, the local scores need to be combined to some score for the form, a global score (unit score). We present a unified global score function. We show that two common choices of global score (the sum of the item scores and the maximum of the item scores) are in fact at either extreme of the feasible region.

3. For the Horvitz-Thompson estimator a common item score is

$$\tilde{\delta}_{kj} = w_k |\tilde{y}_{kj} - z_{kj}| / \varphi_j \quad (1)$$

where w_k is a design weight, \tilde{y}_{kj} is a prediction of the true value y_{kj} , and z_{kj} is the reported value for unit k and variable j (Lawrence and McKenzie 2000). The item score is standardised through some measure φ for variable j , for example an estimate of the total $t_j = \sum_U y_{kj}$ based on values from the previous period of the survey as a standardisation function.

We often take the sum of the item scores as the unit score:

$$g_{sum}(\mathbf{y}_k) = \sum_j \gamma_{kj} \quad (2)$$

where γ_{kj} is an item score for item j , $j = 1, 2, \dots, p$, in record k (e.g. Latouche and Berthelot 1992). For example, the item score γ_{kj} may be the one in (1). Another common choice is the maximum of the item scores as the unit score

$$g_{\max}(\boldsymbol{\gamma}_k) = \max_j(\gamma_{kj}) \quad (3)$$

4. Lawrence and McDavitt (1994) and Hedlin (2003) argue that the maximum function provides some protection against small errors slipping through the editing process. Consider observations of two variables A and B . For example, if τ is some predetermined threshold value, the rule may be to edit a data value of A whenever $g(\gamma_{kA}) \geq \tau$, and similarly for B . Observation k is followed up when $\max(\gamma_{kA}, \gamma_{kB}) \geq \tau$. Alternatively, we may want to follow up k if γ_{kA} and γ_{kB} are both fairly large even if neither of them reaches τ . In this case, the rule may be to edit the form if $\gamma_{kA} + \gamma_{kB}$ oversteps some threshold value. This is the sum function (2).

5. Farwell (2005) proposes a compromise between the sum and the maximum based on the Euclidean distance:

$$g_{esum}(\boldsymbol{\gamma}_k) = \sqrt{\sum_j \gamma_{kj}^2} \quad (4)$$

6. In the next section, the score functions (2) – (4) are unified to become special cases of a general score function. Section III and IV address choice of unit and item score functions. The paper concludes with a discussion in section V.

II. THE UNIT SCORE AS A DISTANCE

A. Distances and geometry

7. Denote a unit score by $g(\boldsymbol{\gamma}_k; \lambda)$, where $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, \dots, \gamma_{kp})'$ with $\gamma_{kj} \geq 0, j = 1, 2, \dots, p$, is a generic notation for the vector of p item scores in record k and λ is a parameter that we will need later on. The unit score functions (2) – (4) are special cases of

$$g(\boldsymbol{\gamma}_k; \lambda) = \left(\sum_{j=1}^p \gamma_{kj}^\lambda \right)^{\lambda^{-1}} \quad (5)$$

where $\lambda \geq 1$, with $g(\boldsymbol{\gamma}_k; 1) = g_{sum}(\boldsymbol{\gamma}_k)$, $g(\boldsymbol{\gamma}_k; 2) = g_{esum}(\boldsymbol{\gamma}_k)$ and $\lim_{\lambda \rightarrow \infty} g(\boldsymbol{\gamma}_k; \lambda) \rightarrow g_{\max}(\boldsymbol{\gamma}_k)$ (e.g. Friedman 1982). The function $g(\boldsymbol{\gamma}_k; \lambda)$ is known as Minkowski's distance or Minkowski's metric. We shall for brevity often omit the parameter λ from (5). We will denote $\lim_{\lambda \rightarrow \infty} g(\boldsymbol{\gamma}_k; \lambda)$ by $g(\boldsymbol{\gamma}_k; \infty)$.

8. It is useful to view all the unit scores as distances in the usual mathematical sense. We view an item score as the distance from the origin $\mathbf{0} = (0, \dots, 0, \dots, 0)$ to the point $(0, \dots, \gamma_{kj}, \dots, 0)$ and a unit score as the distance from $\mathbf{0}$ to $(\gamma_{k1}, \dots, \gamma_{kj}, \dots, \gamma_{kp})$. We refer to $(0, \dots, \gamma_{kj}, \dots, 0)$ as a marginal point of record k . All possible points $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{kj}, \dots, \gamma_{kp})'$ with $\gamma_{kj} \geq 0, j = 1, 2, \dots, p$, constitute together with their distances a metric space.

9. The distance in (4) is the well-known Euclidean distance. The distance that corresponds to the sum function (2) is referred to as the city block distance, Manhattan distance, taxi cab distance, etc; the maximum function (3) is known as the Chebyshev's distance, the chessboard distance, the supremum distance, etc. In statistics, the Euclidean distance is often generalised to the Mahalanobis distance (e.g. Krzanowski 1990). The Mahalanobis distance was used in editing by Hedlin (2003).

10. Hence the unit score can be viewed as a distance also in the usual intuitive sense. In Figure 1 records with $p = 3$ variables and three item scores are envisaged. The Euclidean distance is the distance along the diagonal d . The distance (3) is the longest side, in this case a . The distance (2) is the sum of the sides.

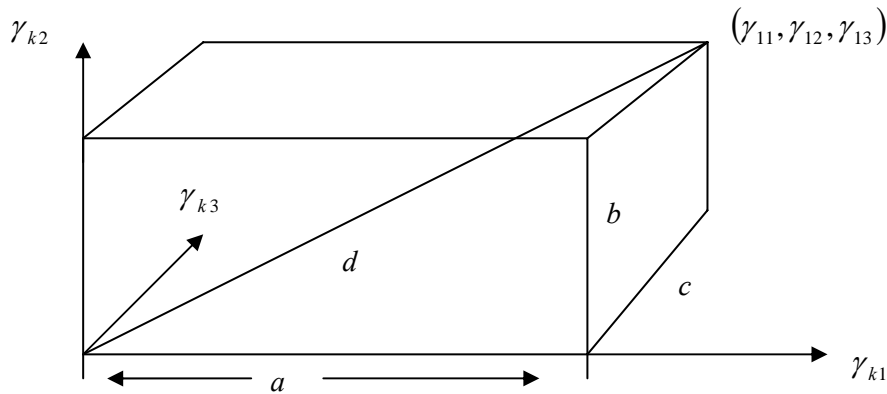


Figure 1. A geometrical interpretation of three types of unit score that summarises three item scores for record $k = 1$. The unit score is the distance from the origin to the point $\boldsymbol{\gamma}_k = (\gamma_{11}, \gamma_{12}, \gamma_{13})$ along three different routes: the shortest distance (d), the longest side (a) and along the edges of the block ($a+b+c$)

11. The points for which $g(\boldsymbol{\gamma}_k)$ is constant form a level surface (level curve if $p = 2$). If it is decided that all records with $g(\boldsymbol{\gamma}_k)$ larger than some threshold τ should be edited, we have the following decision rule:

If $g(\boldsymbol{\gamma}_k) \geq \tau$ then record k will be edited. (6)

12. In other words, the decision rule states that records in a sample s whose unit scores are “inside” the level surface $g(\boldsymbol{\gamma}_k) = \tau$ should remain unedited. In two dimensions the records that will remain unedited under this decision rule form a generalised circle under the particular metric. For the Euclidean metric the level curve $g(\boldsymbol{\gamma}_k; 2) = \tau$ is the natural circle. As all item scores are nonnegative we focus on the first quadrant. The two-dimensional graphs in Figure 2 exhibit the records that will remain unedited under the unit scores (2)-(5). The graphs are ordered from least shaded area to the largest one, that is, by growing λ . Records with item scores falling outside each shaded area will be edited under the metric chosen. To make the graphs comparable, the demarcation of records that should be edited is set at the threshold value $\tau = 0.08$ for all unit score functions.

13. In three dimensions the shaded areas will be the part of an octahedron, a sphere, a superellipsoid and a cube, respectively, that fall in the first octant. (Some of these solids can be viewed at Ron Knott’s website <http://www.mcs.surrey.ac.uk/Personal/R.Knott/Fibonacci/phi3DGeom.html>)

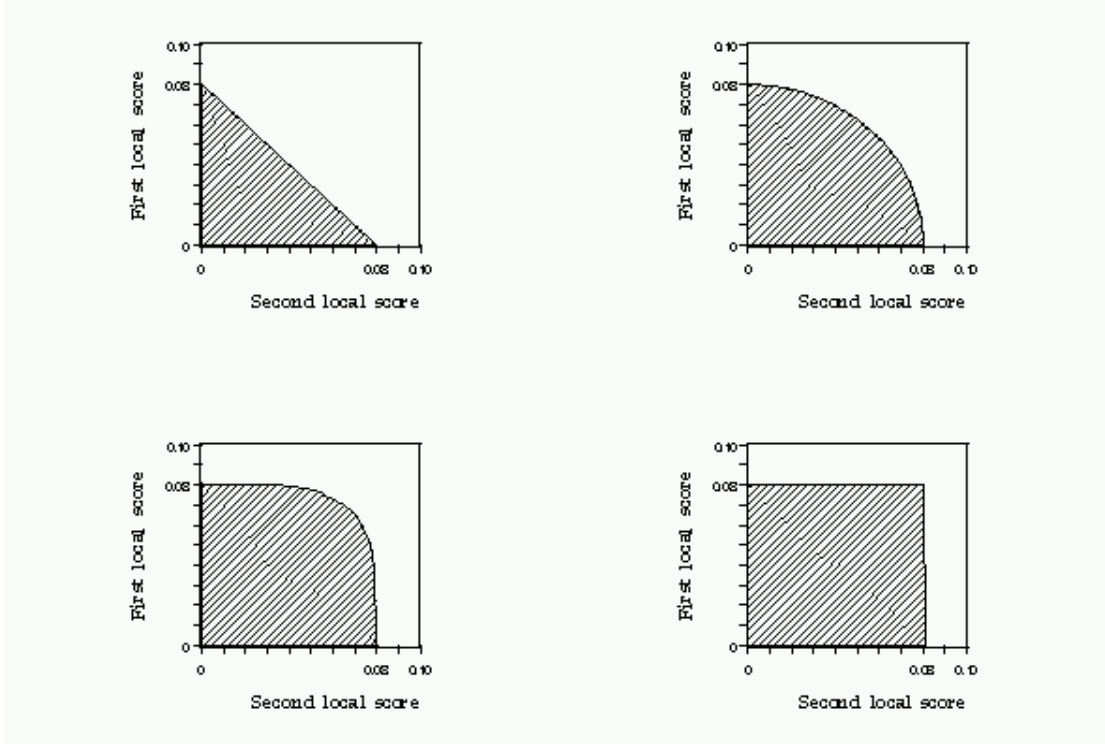


Figure 2. Plots of γ_{k1} against γ_{k2} for a sample $k = 1, 2, \dots, n$. Records with unit scores $g(\gamma_k; \lambda) > 0.08$ fall outside the shaded areas and will be edited. Clockwise starting from top left: the sum function (2), the Euclidean function (4), Minkowski's distance with $g(\gamma_k; 4)$ and the max function (3).

B. Mathematical properties of a unit score

14. A unit score function should have the mathematical properties of a norm to work well for the purposes of editing. The unit score should be a homogeneous, translation invariant metric (distance). We require the following properties; a unit score $g(\gamma_k)$ should

1. be zero if and only if $\gamma_k = \mathbf{0}$
2. be symmetric, i.e. the distance from the origin to γ_k is the same in either direction
3. obey the triangle inequality, i.e. $g(\gamma_k + \gamma_l) \leq g(\gamma_k) + g(\gamma_l) \forall \gamma_k, \gamma_l$
4. be translation invariant, i.e. if the same constant vector is added to the starting and ending point of a item score, the unit score remains the same (we will not change the unit score if all the item scores in Figure 1 are translated away from the origin)
5. be homogeneous, in the sense that $|a|g(\gamma_k) = g(a\gamma_k)$; that is, if all item scores are multiplied by a constant, the unit score is multiplied with the absolute value of the same constant

It follows from properties 1-3 that the unit score is nonnegative, i.e.

$$g(\gamma_k) \geq 0, \forall \gamma_k \quad (7)$$

15. The following demonstrates why the triangle inequality is a necessary property. For simplicity consider the two-dimensional case, that is, records with unit scores $\gamma_k = (\gamma_{kA}, \gamma_{kB})'$. Since $\gamma_k + \mathbf{0} = \gamma_k$, it is true that

$$g(\gamma_k + \mathbf{0}) = g(\gamma_k) \quad (8)$$

Note that properties 1-3 are consistent with (8). Suppose property 3 does not hold. Then there may be a record k for which $\gamma_{kA} > 0$ and $\gamma_{kB} = 0$ such that $g[(\gamma_{kA}, 0) + (0, 0)] > g[(\gamma_{kA}, 0)]$, which is clearly not reasonable.

C. Unit score functions are indexed by λ

16. We explore some further consequences of properties 1-5. First, we will show that the choice of unit scores in Figure 2 is in effect a choice of how much of the striped triangle in Figure 3 should be edited.

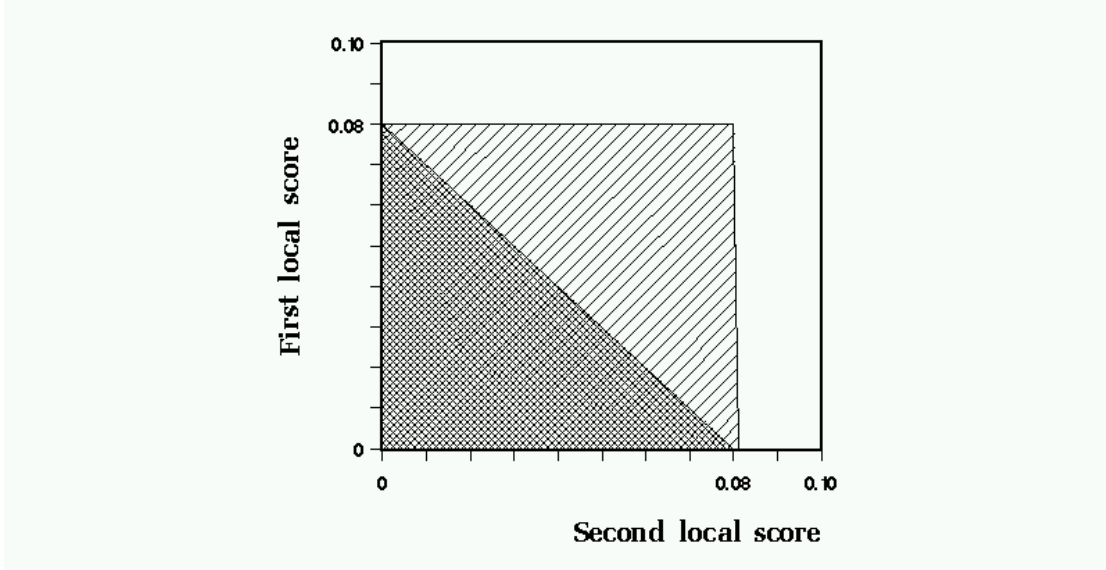


Figure 3. Plots of γ_{k1} against γ_{k2} for a sample $k = 1, 2, \dots, n$. Records with unit scores falling in the shaded area will not be edited no matter choice of unit score function. Records in the striped area may be edited depending on the choice of unit score function, while those in the white area will always be edited

17. It can be shown that for $\lambda < 1$, Minkowski's distance does not satisfy the triangle inequality and hence it is not a metric. The viable range of λ is $\lambda \geq 1$. For $\lambda = 1$, Minkowski's distance always satisfies the triangle inequality with exact equality. As a geometrical interpretation of the level curve $g(\gamma_k; 1) = \tau$ consider the records k on the straight line

$$g[(\gamma_{kA}, 0); 1] = \tau - g[(0, \gamma_{kB}); 1] \quad (9)$$

The graph of the function (9) is the shared hypotenuse of the triangles in Figure 3.

18. Assuming without loss of generality that $\gamma_{kA} > \gamma_{kB}$, we have

$$g(\gamma_k; \infty) = \tau \quad (10)$$

for all k such that $g[(\gamma_{kA}, 0); \infty] = \tau$. This is the function whose graph is the horizontal leg of the striped triangle in Figure 3. Hence this triangle encloses the area which gives a reasonable choice of unit score under Minkowski's distance.

19. Also note that with $g(\gamma_k; 1)$ every item score contributes to the unit score proportionally to its length. With $g(\gamma_k; \infty)$ only one item score contributes to the unit score and hence $g(\gamma_k; \infty)$ is virtually independent of all item scores but one. The parameter λ orders the score function by the equality of impact of item scores from most equal impact to least.

20. Figure 2 may give the impression that using the sum function may lead to an inordinate number of records to edit, since the shaded area is the smallest for this choice of unit score. In practice, however, we often use a larger threshold value τ for the sum function over many variables than for the max function over few variables and hence the workload may actually be less with the sum function.

III. EDITING SITUATIONS, CRITERIA AND FRAMEWORK

21. To be able to discuss choice of unit score function we need to distinguish different editing situations. In terms of error structure, there is measurement bias and measurement variance in the balance. Errors in the data for one business may be highly correlated with those for other businesses. The reason may be questionnaire design that may invoke the same type of error by responders or processing errors on the part of the statistical agency. Some errors may be very large, such as some unit errors (e.g. monetary values given in kronor instead of the requested 1000 kronor) and scanning errors. In principle, there are three editing situations in terms of error structure:

- i. Very large errors remain in data (such as unit errors)
- ii. Few large errors remain but there may nevertheless be non-negligible bias due to many small errors of the same type
- iii. Frequent occurrence of errors of the first two types have largely been cleared out from the processes through continuous improvement; the errors consist now mostly of random measurement errors uncorrelated over observations

The very large errors in Situation *i* are easy to notice and correct. We assume that we are in Situation *iii* and return to the other situation *ii* in section V.

22. We make the following stipulations.

1. If one item on a form is checked, so are all other items on the same form.
2. The cost of checking and editing an item is the same for all items, irrespective of the data value being erroneous or not.
3. All forms contain the same number of items.
4. The measurement model M is $y_{kj} = z_{kj} + \varepsilon_{kj}$, where ε_{kj} is random measurement error associated with each reported data value z_{kj} . Under Situation *iii* we may assume that ε_{kj} and ε_{lj} , $k \neq l$, are uncorrelated.
5. When a data value is edited the result is $y_{kj} = z_{kj}$.

23. There are several desirable aims of editing, including the following criteria.

- 1) Errors left after selective editing should, in some sense, be as small as possible
 - a) Minimum Mean Squared Error (MSE) under fixed cost
 - b) Least bias for each variable under fixed cost
- 2) The editing process should allow the producer to keep the effect of errors under control
 - a) Fixed maximum MSE for each variable
 - b) Fixed maximum bias

In Situation *iii* we focus on minimum measurement variance rather than minimum MSE or bias.

24. We discuss item and unit score functions under these stipulations and criteria. For a wider discussion of purposes of editing, see for example Granquist and Kovar (1997).

IV. WHICH UNIT SCORE FUNCTION?

A. Errors left after selective editing should be as small as possible

25. We shall examine what unit score function is suitable for a minimum MSE criterion. To keep notation simple we shall for the moment consider only the univariate case. The measurement variance of the estimator \hat{t}_{ys} conditional on the sample s is

$$Var_M(\hat{t}_{ys}) = Var_M\left(\sum_{k \in s} \hat{z}_k + \hat{\varepsilon}_k\right) = \sum_{k \in s} w_k^2 \sigma_k^2 \quad (11)$$

where M refers to the measurement error model, $\sigma_k^2 = Var_M(\varepsilon_k)$ and π_k is the inclusion probability with $w_k = \pi_k^{-1}$ being a design weight. If the set r is edited, $\sigma_k^2 = 0$ and $\varepsilon_k = 0$ for $k \in r$. In Situation *iii* the bias is negligible and we can make the approximation

$$\Delta MSE(\hat{t}_{ys}) = MSE_M(\hat{t}_{ys}) - MSE_M(\hat{t}_{ysr}) \approx \sum_{k \in s-r} w_k^2 \sigma_k^2 \quad (12)$$

where \hat{t}_{ysr} is the estimate based on sample s with r edited. The set r that maximises $\Delta MSE(\hat{t}_{ys})$ is the set of records with the largest $w_k^2 \sigma_k^2$. A decision rule is then under (12):

$$\text{if } w_k^2 \sigma_k^2 \geq \tau_{MSE}(1) \text{ then the item in record } k \text{ will be edited} \quad (13)$$

where the parameter $p = 1$ indicates the univariate case. In the multivariate case we have $\Delta MSE(\hat{t}_{y_j})$, $j = 1, 2, \dots, p$. To minimise the sum of the $\Delta MSE(\hat{t}_{y_j})$, note that the set r that yields

$$\max_r \sum_{k \in r} \sum_{j \in k} \Delta MSE(\hat{t}_{y_j}) \text{ is the set containing the records with the largest } w_k^2 \sum_j \sigma_{kj}^2. \text{ The decision rule is}$$

then

$$\text{if } w_k^2 \sum_j \sigma_{kj}^2 \geq \tau_{MSE}(p) \text{ then the item in record } k \text{ will be edited} \quad (14)$$

26. For this to be useful there must be a simple way of approximating $Var(\varepsilon_{kj}) = \sigma_{kj}^2$. For the purposes of selective editing, it is reasonable to assume that

$$Var(\varepsilon_{kj}) \propto (\tilde{y}_{kj} - z_{kj})^2. \quad (15)$$

Hence with $\tilde{\delta}_{kj}$ defined as in (1) the criterion of minimum MSE under fixed cost leads to the Euclidean

unit score with $\gamma_{kj} = (\tilde{\delta}_{k1}, \tilde{\delta}_{k2}, \dots, \tilde{\delta}_{kp})'$ and $\lambda = 2$.

B. Be in control

27. The rule in (14) does not guarantee that the error for a particular variable is within some bound. For this a rule that treats each variable separately is required. As we have seen in section IIC it is necessary to use the max function to meet this aim. A decision rule based on this criterion is

$$\text{if } \max_j (\gamma_{kj}) \geq \tau \quad (16)$$

The decision rule in (16) cannot operate under a fixed budget, at least not if the budget constraint is imposed strictly, as the number of records to be edited cannot be determined ahead of time.

V. DISCUSSION

28. In this paper we contend that common unit score functions share the same form which can be expressed as Minkowski's distance function with the sum function and the max function as the two extreme choices. This puts the various unit score functions in the same basket and shows how they are connected. Also, it makes it easy to implement all unit score functions in a software. Putting $\lambda = 10$ for the max function is adequate for most practical purposes.

29. We have discussed the best choice of unit score function in a situation where errors are mostly random and symmetric. We have argued that in this situation either the Euclidean unit score function proposed by Farwell (2005) or the max function is a good combination of the item scores in (1). If minimum overall error under a fixed budget is desired, the Euclidean unit score function is to be preferred. The Euclidean unit score function does not impose a limit for the bias of a particular variable. If such a limit is called for, it is necessary to make use of the maximum unit score function.

30. We will briefly discuss Situation *ii*. In this situation we want to look at not only (12) but also at the decrease in bias due to editing: $\sum_{k \in s-r} \sum_{j \in k} \Delta \text{Bias}(\hat{t}_{y_j})$.

Usually $\sum_{k \in r} \sum_{j \in k} \Delta \text{Bias}(\hat{t}_{y_j})$ will be maximised by the set containing the records with the largest $\sum_j |\hat{\varepsilon}_{kj}|$.

With the approximation $|\varepsilon_{kj}| \propto |\tilde{y}_{kj} - z_{kj}|$ the global score function is the sum function with

$\gamma_{kj} = (\tilde{\delta}_{k1}, \tilde{\delta}_{k2}, \dots, \tilde{\delta}_{kp})'$ and $\lambda = 1$. However, to see that there is no guarantee that editing under a fixed budget will improve estimates in terms of bias, consider the six records to be subjected to selective editing displayed in Table 1. If no record is edited the error $\hat{t}_j - t_j = \sum_s \hat{\varepsilon}_{kj}$ is zero. If the total cost of editing is fixed so that m records can be edited, the error $\hat{t}_j - t_j = \sum_s \hat{\varepsilon}_{kj}$ will be larger than 0 unless $m = 6$ or $m = 0$.

Table 1. A sample s consisting of six observations.

Observation	$\hat{\varepsilon}_{kj}$
1	-1
2	-1
3	-1
4	-1
5	-1
6	5

References

- Farwell, K. (2005). Significance Editing for a Variety of Survey Situations. Paper presented at the 55th session of the International Statistical Institute, Sydney, 5–12 April.
- Friedman, A. (1982). Foundations of Modern Analysis. New York: Dover.
- Granquist, L. and Kovar, J.G. (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality, eds L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: Wiley, 415-435.
- Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the UK Office for National Statistics. Journal of Official Statistics, 19, 177-199.
- Krzanowski, W.J. (1990). Principles of Multivariate Analysis. A User's Perspective. Oxford Science Publications.
- Latouche, M. and Berthelot, J.M. (1992). Use of a Score Function to Prioritise and Limit Recontacts in Business Surveys. Journal of Official Statistics, 8, 389-400.
- Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. Journal of Official Statistics, 10, 437-447.
- Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243-253.
