# Theory of selective editing with score functions

Dan Hedlin
Statistics Sweden
Box 24300; SE-10451 Stockholm, Sweden
dan.hedlin@scb.se

## 1. Unit scores in selective editing

Errors in data haunt practitioners in statistics. It is sometimes possible to recontact the individual or, in general terms, the object, and make another observation. It is usually necessary to prioritize those objects that are most cost-effective to recontact. For multivariate observations there are several items per unit. Usually we want to edit and verify all observed values of the unit in one go rather than each item separately, that is, we want to select a unit to recontact, not an item. So the item scores need to be combined to a *unit score*. The unit score indicates the importance of recontacting the unit.

We need to distinguish different editing situations. In terms of error structure, there is measurement bias and measurement variance in the balance. Some errors may be very large, such as scanning and other recording errors and misreporting errors including unit errors (e.g. monetary values given in SEK instead of the requested SEK1000). In principle, there are three *editing situations* in terms of error structure:

i. Very large errors in the data (such as unit errors)
ii. No very large errors remain but there may nevertheless be non-negligible bias due to many small errors of the same type
iii. Errors of the first two types have largely been cleared out through continuous improvement of measurement processes; the errors consist now mostly of zero-mean random measurement errors uncorrelated over observations from different units. Items within unit may be correlated.

The very large errors in Situation *i* should not be difficult to identify. Errors that arise in Situation *ii* are sometimes referred to as inliers (Granquist 1995). In Situation *iii* there may under some circumstances be little need for editing. Hedlin (2003, p. 193) discusses in what situations the practitioner can rely on the law of large numbers to cancel out errors when computing estimates. Errors in data for one business may be correlated with those for other businesses. The reason may be questionnaire design that invokes the same type of error by responders or systematic processing errors on the part of the statistical agency. That is to say, it is not uncommon for business surveys to be in Situation *ii*.

While the item score (1) introduced below is motivated from a design-based survey perspective centring on a particular estimator, we shall in Section 2 show that there is a more fundamental reason why $\left|\tilde{y}_{kj} - z_{kj}\right|$, where $\tilde{y}_{kj}$ is a prediction of $y_{kj}$, is a generally

useful building block. Consider the Horvitz-Thompson (HT) estimator $\hat{t}_j = \sum_s w_k z_{kj}$ of the total $t_j = \sum_U y_{kj}$, where *s* denotes the sample and *U* the target population. The $w_k$ 's are survey weights. In the univariate case we shall often drop the subscript *j*.

Let an item score be $\left|\hat{t}_{(k)} - \hat{t}\right|$, where $\hat{t}_{(k)}$ is the same as $\hat{t}$ except that for unit *k* the observed value of the study variable is replaced by the predicted value. Then the item score for the HT estimator is

$$\tilde{\delta}_{kj} = w_k \left|\tilde{y}_{kj} - z_{kj}\right|, \qquad (1)$$

(Latouche and Berthelot 1992, Lawrence and McDavitt 1994). The unit score is often used to divide data into two lots: if the unit score is above a threshold, observations of the unit's values should be edited. If it is below, the unit is either left unattended or some automatic editing is performed. Alternatively, a random subsample may be drawn with inclusion probability proportional to unit score. Optimal subsample design will be treated elsewhere. Even if the ambition is to follow up all units, the unit score will give management a tool for prioritization of the order of the work.

Denote a unit score function by $g(\boldsymbol{\gamma}_k)$, where $\boldsymbol{\gamma}_k = (\gamma_{k1}, ..., \gamma_{kp})'$ with $\gamma_{kj} \geq 0$, $j = 1, 2, \ldots, p$, is a generic notation for the *p*-vector of item scores for unit *k*. Latouche's and Berthelot's (1992) unit score is $g_{sum}(\boldsymbol{\gamma}_k) = \sum_j \gamma_{kj}$, whereas Lawrence and McDavitt (1994) and Hedlin (2003) use the maximum of the item scores as their unit score, $g_{\max}(\boldsymbol{\gamma}_k) = \max_j(\gamma_{kj})$. Farwell (2005) proposes a compromise between the sum and the maximum based on the Euclidean distance: $g_{esum}(\boldsymbol{\gamma}_k) = \sqrt{\sum_j \gamma_{kj}^2}$.

The unit score functions $g_{sum}$, $g_{esum}$ and $g_{max}$, referred to as the sum, the Euclidean and the max function, respectively, are special cases of

$$g(\boldsymbol{\gamma}_k; \lambda; p) = \left( \sum_{j=1}^{p} \gamma_{kj}^{\lambda} \right)^{\lambda^{-1}} \qquad (2)$$

where $\lambda \geq 1$, with $g(\boldsymbol{\gamma}_k; 1; p) = g_{sum}(\boldsymbol{\gamma}_k)$, $g(\boldsymbol{\gamma}_k; 2; p) = g_{esum}(\boldsymbol{\gamma}_k)$ and $\lim_{\lambda \to \infty} g(\boldsymbol{\gamma}_k; \lambda; p) \to g_{\max}(\boldsymbol{\gamma}_k)$ (e.g. Friedman 1982). The function (2) is known as Minkowski's distance or metric.

## 2. Which unit and item score functions?

We shall discuss various unit score functions under the following stipulations.

1. If one item in a unit is edited, so are all other items for the same unit.
2. The cost of editing an item is the same for all items, irrespective of the data value being erroneous or not.
3. The measurement model *M* is $Y_{kj} = Z_{kj} + R_{kj}$, where $R_{kj}$ is a measurement error associated with the reported data value $Z_{kj}$.

4. In Situation *iii*, $\text{cov}_M\left(R_{ki}, R_{lj}\right) = 0$ and $E_M\left(R_{ki}\right) = 0$, $\forall i, j$ and $k \neq l$, where $M$ refers to the measurement error model.
5. When a data value is edited the result is $y_{kj}$.

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

1) Errors remaining after selective editing should, in some sense, be as small as possible:
   a) In Situation *iii*, minimum measurement variance under fixed cost
   b) In Situation *ii*, minimum bias in absolute terms under fixed cost
2) The editing process should allow the producer to control the effect of errors:
   a) Fixed maximum measurement variance for each variable
   b) Fixed maximum bias in absolute terms for each variable

A common criterion for biased estimators is minimum MSE, which is an alternative to Aim 1a. In official statistics and many other applications, however, non-trivial bias is undesirable, even if the MSE is low.

We discuss unit score functions under these stipulations and criteria. We shall examine what unit score functions are suitable for Aims 1 a-b and 2 a-b. To keep notation simple we shall for the moment consider only the univariate case. Let the HT-estimator $\hat{t}_y$ with the subsample $a \subset s$ edited be denoted by $\hat{t}_{y;a}$. The measurement variance of the error in $\hat{t}_{y;\varnothing}$ conditional on *s* is

$$\text{var}_M\left(\hat{t}_{y;s} - \hat{t}_{y;\varnothing}\right) = \text{var}_M\left(\sum_s w_k R_k\right) = \sum\sum_s w_k w_l \,\text{cov}_M\left(R_k, R_l\right)$$

If the set *a* is edited to meet Aim 1a then under Stipulation 5,

$$\text{var}_M\left(\hat{t}_{y;s} - \hat{t}_{y;a}\right) = \sum_{s-a} w_k^2 \sigma_k^2 ,$$

where $\sigma_k^2 = \text{var}_M\left(R_k\right)$. The set *a* containing the units with the largest $w_k^2\sigma_k^2$ minimizes $\text{var}_M\left(\hat{t}_{y;s} - \hat{t}_{y;a}\right)$. To minimize the sum of the measurement variances in the multivariate case, $\sum_j \text{var}_M\left(\hat{t}_{y_j;s} - \hat{t}_{y_j;a}\right)$, the set *a* will be the set containing the units with the largest $v_{k\bullet} = \sum_j v_{kj} = w_k^2 \sum_j \sigma_{kj}^2$ .

Proneness to measurement error often varies from respondent to respondent. If $y_{kj}$ were known ahead of recontact, it would be reasonable to estimate $\sigma_{kj}$ in Situation *iii* with $c\left|y_{kj} - z_{kj}\right|$ for some constant $c > 0$. To estimate $\sigma_{kj}^2$ in Situation *iii* we use

$$\hat{v}_{kj} = w_k^2 \hat{\sigma}_{kj}^2 = w_k^2 \left(\tilde{y}_{kj} - z_{kj}\right)^2 \qquad (3)$$

disregarding the constant *c* which will not alter the order of observations given by the unit score. Hence with $\tilde{\delta}_{kj}$ defined as in (1) the criterion of minimum $\sum_j \text{var}_M\left(\hat{t}_{y_j;s} - \hat{t}_{y_j;a}\right)$ under fixed cost in Situation *iii* leads to the Euclidean unit score with $\boldsymbol{\gamma}_k = \left(\tilde{\delta}_{k1}, ..., \tilde{\delta}_{kp}\right)'$ and $\lambda = 2$. Note that we in (3) recover the building block

$\left|\tilde{y}_{kj} - z_{kj}\right|$ in (1). With many other estimators of other parameters the quantity in (3) would still be central although the weights $w_k$ would be different.

Turning to Situation *ii* and Aim 1b, consider

$$\sum_j \left| bias_M\left(\hat{t}_{y_j;a}\right)\right| = \sum_j \left|\sum_{s-a} w_k \mu_{kj}\right| \qquad (4)$$

where $\mu_{kj} = E_M\left(R_{kj}\right)$. An upper bound of (4) is $\sum_j \sum_{s-a} \left|w_k \mu_{kj}\right|$, which is minimized by the set $a$ containing the units with the largest $b_{k\bullet} = \sum_j \left|w_k \mu_{kj}\right|$. As an estimator of $b_{kj} = \left|w_k \mu_{kj}\right|$ we may use (1). Hence with $\tilde{\delta}_{kj}$ defined as in (1) the criterion of minimum $\sum_j \sum_{s-a} \left|w_k \mu_{kj}\right|$ under fixed cost and Situation *ii* the unit score function is the sum function with $\boldsymbol{\gamma}_k = \left(\tilde{\delta}_{k1}, \tilde{\delta}_{k2}, ..., \tilde{\delta}_{kp}\right)'$ and $\lambda = 1$. Minimizing mean squared error calls for a compromise between $\lambda = 1$ and $\lambda = 2$.

However, to see that editing under a fixed budget will not necessarily improve estimates in terms of bias, consider a sample $s$ of six units with realized values $w_k R_k = w_k r_k = -1$ for $k = 1, 2, ..., 5$, and $w_6 r_6 = 5$. If no unit is edited the realized error in $\hat{t}_{y;\varnothing}$, that is $\sum_s w_k r_k$, vanishes. If we can afford having a set $a$ edited, the realized error $\sum_{s-a} w_k r_k$ will be larger than zero unless $a = s$ or $a = \varnothing$.

Consider now Aims 2a-b. It should be noted that neither editing strategy so far guarantees that the error for a particular variable is within some bound. The max function is the only unit score that meets this requirement. Suppose we need an $a$ that makes either $\sum_a w_k^2 \sigma_{kj}^2$ or $\sum_a w_k \mu_{kj}$ smaller than some predetermined number *for each j*. This strategy cannot operate under a fixed budget, at least not if the budget constraint is imposed strictly. One way to control $\sum_a w_k^2 \sigma_{kj}^2$ (or $\sum_a w_k \mu_{kj}$) for a particular $j$ is to send to editing any unit where $\hat{v}_{kj}^2 \geq \tau$ (or $w_k \left|\tilde{y}_{kj} - z_{kj}\right| \geq \tau$) for any $j = 1, 2, \ldots, p$. The unit score function is the max function in both cases.

A potential problem with (1) is the prediction error in $\tilde{y}_{kj}$. It can be shown that the max function is more robust than unit score functions with finite $\lambda$.

# 3. Application

A general issue with the evaluation of real edited data is that the errors documented are only the errors found. To gain control of the errors we constructed a data set with true values $\theta_{kj}$, $j = 1, 2, 3$, $k = 1, \ldots, 10000$, in accordance with the gamma distributed Populations 1, 3 and 12 in Lee et al. (1994). For each unit a true, an erroneous and a predicted value were generated. An erroneous value is the true value with a

measurement error added, which may be minute. 29% of the reported units were erroneous. In an erroneous unit, all values were incorrect. They were generated as $z_{k1} = \theta_{k1} + \exp(x_{k1})$, $z_{k2} = \theta_{k2} + \exp(z_{k1}^{0.1} x_{k2})$ and $z_{k3} = \theta_{k3} + \theta_{k3}^{0.5} x_{k3}$, where the $x_{kj}$'s were realized pseudo-random standard normal variables. Note that $z_{k1}$ and $z_{k2}$ are correlated within unit. Predicted values were generated as $\tilde{y}_{k1} = 0.96\theta_{k1} + \theta_{k1}^{1.5} x'_{k1}$, $\tilde{y}_{k2} = 0.91\theta_{k2} + \theta_{k2}^{1.5} x'_{k2}$ and $\tilde{y}_{k3} = 0.89\theta_{k3} + 0.2\theta_{k3}^{0.25} x'_{k3}$, with the $x'_{kj}$'s taken from a standard normal distribution. The predicted values are meant to resemble values from a previous wave of the survey. Then 1000 populations were created with new reported values in each population. The true and predicted values were retained over simulations. A second and a third suite of 1000 populations each were created. The erroneous values were generated as $z_{k1} = \theta_{k1} + 3\exp(\theta_{k1}^{0.2} x''_{k1})$, $z_{k2} = \theta_{k2} + 5\exp(\theta_{k2}^{0.1} x''_{k2})$ and $z_{k3} = \theta_{k3} + \theta_{k3}^{0.5} x''_{k3}$ for suite 2 and $z_{k1} = \theta_{k1} + \theta_{k1}^{0.25} x'''_{k1}$, $z_{k2} = \theta_{k2} + \theta_{k2}^{0.1} x'''_{k2}$ and $z_{k3} = \theta_{k3} + \theta_{k3}^{0.5} x'''_{k3}$ for suite 3. Note that suites 1 and 2 are in Situation *ii*. The predicted values were generated through the same formulae in all suites. We refer to the editing methods that unrealistically use the true value instead of the predicted value to 'ideal methods'.

The sum, Euclidean and max unit score functions, both the realistic and ideal versions, were applied to the populations with 200 units edited. When a value was edited the true value was recovered. Biases and variances were computed for each variable and unit score. The bias was estimated as $\hat{b}_M(t_{y_j;a}) = \sum_{\iota}^{I} (t_{y_j;\varnothing} - t_{y_j;a})(NI)^{-1}$, $j = 1, 2, 3$, where $t_{y_j;a}$ is the population total in a self-representing sample with set $a$ edited, $N = 10000$ and $I = 1000$ is the number of simulations. The variance was estimated as $\hat{v}_M(t_{y_j;a}) = \sum_{\iota}^{I} \left[t_{y_j;\varnothing} - t_{y_{hj};a} - \hat{b}_M(t_{y_j;a})\right]^2 / I(N-1)$. Table 1 reports on the sum of the biases in absolute terms, $\sum_j \left|\hat{b}_M(t_{y_j;a})\right|$, and the sum of the variances, $\sum_j \hat{v}_M(t_{y_j;a})$, for selective editing with each of three unit score functions and for no editing at all. The potential of selective editing to substantially limit editing is again corroborated. The ideal scores in Table 1 confirm that the Euclidean function is in theory the best choice for minimizing the measurement variance. However, as seen from the realistic scores the max function works better in practice in two out of three suites due to its greater robustness to prediction error. The greatest difference between unit score functions should be found in data with positive correlation within units. Indeed, for suite 1 the sum function performs relatively worse than for suites 2 and 3. We may also suspect that the difference between the sum function and the others grows with number of variables and with number of units edited up to a limit where the editing becomes close to being exhaustive. This is borne out by simulations not shown here.

In terms of bias there is little difference between the realistic unit score functions although the sum function is the best one among the ideal score functions for data without extreme errors. For suite 3, which has normally distributed errors uncorrelated

within unit, there is little to choose between in terms of unit score. Note that the editing through realistic unit scores increases the bias in suite 3.

**Table 1.** Sum of variances and biases under six unit scores and no editing

| | Sum of variances | | | Sum of biases in absolute terms | | |
|---|---|---|---|---|---|---|
| | Suite 1 | Suite 2 | Suite 3 | Suite 1 | Suite 2 | Suite 3 |
| Sum function | 34.2 | 868.8 | 26.9 | 1.25 | 8.67 | 0.0807 |
| Euclidean function | 33.1 | 862.9 | 26.4 | 1.23 | 8.67 | 0.0924 |
| Max function | 32.5 | 863.2 | 26.0 | 1.22 | 8.67 | 0.0995 |
| Ideal sum function | 30.6 | 865.4 | 24.1 | 1.10 | 8.66 | 0.0015 |
| Ideal Euclidean function | 30.3 | 861.7 | 24.0 | 1.12 | 8.66 | 0.0015 |
| Ideal max function | 30.4 | 862.0 | 24.0 | 1.12 | 8.67 | 0.0016 |
| No editing | 57.7 | $2.9 \cdot 10^6$ | 29.2 | 1.36 | 20.59 | 0.0022 |

# 4. Discussion

We have shown that unit score functions widely used in surveys share the same form which can be expressed as Minkowski's metric 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 related. It facilitates software implementation.

We have discussed the best choice of unit score function in a situation where errors in different units are uncorrelated and have zero mean. We have argued that in this situation either the Euclidean unit score function proposed by Farwell (2005) or the maximum unit score function is a good combination of the item scores in (1) with a strong preference of the latter due to its greater robustness. 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.

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. (1995). Improving the Traditional Editing Process. In *Business Survey Methods*, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott, New York: Wiley, 385-401.

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.

LATOUCHE, M. and BERTHELOT, J.M. (1992). Use of a Score Function to Prioritize 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.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics*, 10, 231-243.