Correcting data violating linear restrictions using the deducorrect and editrules packages

Mark van der Loo[1,⋆], Edwin de Jonge[1] and Sander Scholtus[1]

1. Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands
⋆Contact author: m.vanderloo@cbs.nl

Keywords: Data editing, official statistics, linear restrictions, errror correction

editrules In many computational and statistical problems one needs to represent a set of m linear restrictions on a data record x of the form $Ax-b=0$ or $Ax-b\geq 0$,where A is a matrix with real coefficients and b a vector in $\mathbb{R}^m$. Constructing and maintaining A manually is tedious and prone to errors. Moreover, in many cases the restrictions are stated verbosely, for example as "profit + cost must equal turnover". The **editrules** package can parse restrictions written in *R* language to matrix form. For example:

```
> editmatrix(c("x + 3*y == -z", "x>0"))

Edit matrix:
   x y z CONSTANT
e1 1 3 1        0
e2 1 0 0        0

Edit rules:
e1 : x + 3*y == -z
e2 : x > 0
```

The result is an `S3` object of class `editmatrix` which extends the standard `matrix` object. Here, the `editmatrix` function accepts linear restrictions in `character` or `data.frame` format. The latter offers the opportunity to name and comment the restrictions. The package also offers functionality to check data against the imposed restrictions and summarize errors in a useful way. In fact, the error checking functionality is independent of restrictions being of linear form, and can be used for any restriction including numerical and/or categorical data.

deducorrect Raw survey data is often plagued with errors which need to be solved before one can proceed with statistical analysis. The **deducorrect** package offers functionality to detect and correct typing errors (based on the Damerau-Levenshtein distance) and rounding errors in numerical data under linear restrictions. It also solves sign errors and value swaps, possibly masked by rounding errors. The methods used are (slight) generalizations of the methods described by Scholtus (2008) and Scholtus (2009). All datacorrection functions return corrected data where possible, a log of the applied corrections and the correction status. The package also offers functionality to determine if a matrix is totally unimodular, which is useful for solving errors in data involving balance accounts.

References

Scholtus, S. (2008). Algorithms for correcting obvious inconsistencies and rounding errors in business data. Technical Report 08015, Statistics Netherlands, Den Haag. *Accepted by J.Official Stat.*

Scholtus, S. (2009). Automatic correction of simple typing error in numerical data with balance edits. Technical Report 09046, Statistics Netherlands, Den Haag. *Accepted by J. Offical Stat.*