# Perform a statistical adjustement (reweighting)

*Dominique Emmanuel*

*2016-03-01*

## Theory

Let $X^1, \dots, X^n$ be some categorical variables, and for each variable $X^i$ let $m^{i1}, \dots, m^{in_i}$ be its levels, and let $X^{i1}, \dots, X^{in_i}$ be the associated dummies. Let $w$ be an initial weight. We call **adjustement** a weight $w'$ such that :

- $w'$ is as close as possible to $w$ (L2 norm)
- $\sum_{k=1}^{N} w'_k = \sum_{k=1}^{N} w_k$
- $\forall k \;\; 0 \leq w_{\min} \leq w'_k \leq w_{\max}$
- the weighted values of each level $m^{ij}$ is equal (and/or greater and/or lower) to specifed value $v^{ij}$.

Consequently

$$w' = \underset{\substack{\forall i,j \;\; \sum_{k=1}^{N} x_k X_k^{ij} = v^{ij} \\ \sum_{k=1}^{N} x_k = \sum_{k=1}^{N} w_k \\ \forall k \;\; w_{\min} \leq x_k \leq w_{\max}}}{\arg\min} \|x - w\|^2$$

NB: Any equality $\sum_{k=1}^{N} x_k X_k^{ij} = v^{ij}$ can be replaced by an inequality.

## Practice

The function `adjustement` allows to perform this optimisation.

Let's take a subest of `esoph` of 50 individuals :

```
set.seed(123)
data <- esoph[sample(seq(nrow(esoph)),50), ]
w_initial <- rep(nrow(esoph)/nrow(data), nrow(data))
```

Let's define somme margins:

```
table(esoph$agegp)/nrow(esoph)
```

```
##
##     25-34     35-44     45-54     55-64     65-74       75+
## 0.1704545 0.1704545 0.1818182 0.1818182 0.1704545 0.1250000
```

```
table(esoph$alcgp)/nrow(esoph)
```

```
##
## 0-39g/day     40-79    80-119      120+
## 0.2613636 0.2613636 0.2386364 0.2386364
```

```
margins <- list(
  list(var_name = "agegp",
       value = c("25-34" = 0.17, "35-44" = 0.17, "45-54" = 0.17, "55-64" = 0.17),
       min = c("65-74" = 0.17, "75+" = 0.12)
  ),
  list(var_name = "alcgp",
       value = c("0-39g/day" = 0.26),
       min = c("40-79" = 0.3),
       max = c("80-119" = 0.2)
  )
)
```

Let's perform the adjustement:

```
library(adjustment)
adj <- adjustment(data = data, margins = margins, weight = w_initial, weight_min = 0.1, weight_max = 30)
```

```
## Note: method with signature 'diagonalMatrix#sparseMatrix' chosen for function 'rbind2',
##  target signature 'ddiMatrix#ddiMatrix'.
##  "sparseMatrix#diagonalMatrix" would also be valid
```

```
adj$IsError
```

```
## [1] FALSE
```

```
w <- adj$w
```

Let's verify:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data$w <- w
data %>%
  group_by(agegp) %>%
  summarise(n = sum(w)) %>%
  merge(data %>% summarise(n0 = sum(w))) %>%
  mutate(mean = n/n0)
```

```
##   agegp     n n0 mean
## 1 25-34 14.96 88 0.17
## 2 35-44 14.96 88 0.17
## 3 45-54 14.96 88 0.17
## 4 55-64 14.96 88 0.17
## 5 65-74 14.96 88 0.17
## 6   75+ 13.20 88 0.15
```

```
data %>%
  group_by(alcgp) %>%
  summarise(n = sum(w)) %>%
  merge(data %>% summarise(n0 = sum(w))) %>%
  mutate(mean = n/n0)
```

```
##       alcgp        n n0      mean
## 1 0-39g/day 22.88000 88 0.2600000
## 2     40-79 29.34954 88 0.3335175
## 3    80-119 17.60000 88 0.2000000
## 4      120+ 18.17046 88 0.2064825
```

```
hist(w, 10, col = "steelblue", xlab = "new weight",border="white")
```