Statistical Tools for Analysis of Non-Probability samples: Part 2

Jae Kwang kim and Yonghyun Kwon

2025-07-24

1 Recap: Generalized entropy calibration(GEC)

- We maximize the generalized entropy

$$Q_G(\boldsymbol{\omega}) = -\sum_{i \in S} G(\omega_i) \tag{1}$$

subject to

$$\sum_{i \in S} \omega_i \boldsymbol{x}_i = \sum_{i=1}^{N} \boldsymbol{x}_i, \tag{2}$$

where $G(\cdot) : \mathcal{V} \to \mathbb{R}$ is a strictly convex and differentiable function.

- Using the Lagrange multiplier method, we find the minimizer of

$$\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\lambda}) = \sum_{i \in S} G(\omega_i) - \boldsymbol{\lambda}^\top \left(\sum_{i \in S} \omega_i \boldsymbol{x}_i - \sum_{i=1}^{N} \boldsymbol{x}_i \right) \tag{3}$$

with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$.

- By setting $\partial \mathcal{L} / \partial \omega_i = 0$ and solving for ω_i, we obtain

$$\hat{\omega}_i(\boldsymbol{\lambda}) = g^{-1}\left(\boldsymbol{\lambda}^\top \boldsymbol{x}_i \right),$$

where $g(\omega) = dG(\omega)/d\omega$.

- Thus, by plugging $\hat{\omega}_i(\boldsymbol{\lambda})$ into $\mathcal{L}$ in (3), we obtain

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\lambda} \left[\sum_{i \in S} G\left\{ \hat{\omega}_i(\boldsymbol{\lambda}) \right\} - \boldsymbol{\lambda}^\top \left(\sum_{i \in S} \hat{\omega}_i(\boldsymbol{\lambda}) \boldsymbol{x}_i - \sum_{i=1}^{N} \boldsymbol{x}_i \right) \right].$$

1.1 Toy example

Generalized Entropy	$G(\omega)$	$\rho(\nu)$
Squared loss	$\omega^2/2$	$\nu^2/2$
Kullback-Leibler	$\omega\log(\omega)$	$\exp(\nu-1)$
Shifted KL	$(\omega-1)\{\log(\omega-1)-1\}$	$\nu+\exp(\nu)$
Empirical likelihood	$-\log(\omega)$	$-1-\log(-\nu)$
Squared Hellinger	$(\sqrt{\omega}-1)^2$	$\nu/(\nu-1)$
Rényi entropy $(\alpha \neq 0, -1)$	$\frac{1}{\alpha+1}\omega^{\alpha+1}$	$\frac{\alpha}{\alpha+1}\nu^{\frac{\alpha+1}{\alpha}}$

Table 1: Examples of generalized entropies, $G(\omega)$, and the corresponding convex conjugate functions, $\rho(\nu)$

```r
# install.packages(GECal)
library(GECal)
# Sampled study variable
y=c(5, 4, 7, 9, 11, 10, 13, 12, 15, 15)
# Sampled auxiliary variables
Xs=cbind(
  c(1,1,1,1,1,1,1,1,1,1),
  c(1,1,1,1,1,0,0,0,0,0),
  c(1,3,5,7,9,6,7,8,9,10)
)
# vector of population totals
total=c(160,124,700)
# base weights before calibration
d = rep(1, 10)
```

The structure of data is as follows:

i	δ_i	x_{i0}	x_{i1}	x_{i2}	y_i
1	1	1	1	1	5
2	1	1	1	3	4
3	1	1	1	5	7
4	1	1	1	7	9
5	1	1	1	9	11
6	1	1	0	6	10
7	1	1	0	7	13
8	1	1	0	8	12
9	1	1	0	9	15
10	1	1	0	10	15
11	0	NA	NA	NA	NA
$\vdots$	$\vdots$	NA	NA	NA	NA
160	0	NA	NA	NA	NA
Total	10	160	124	700	$\theta=?$

1.1.1 Kullback-Leibler(Exponential tiliting)

```r
# GEC estimator using ET(exponential tilting) divergence
cal_ET <- GEcalib(~ 0 + Xs, dweight = d, const = total,
           method = "GEC0", entropy = "ET")
head(cal_ET$w)
```

```
##         1         2         3         4         5         6
## 48.359404 31.828847 20.948884 13.787987  9.074879 10.475456
```

```r
GECal::estimate(y ~ 1, calibration = cal_ET)$estimate
```

```
##   Estimate Std. Error
## y 1189.612   84.30957
```

1.1.2 Empirical Likelihood

```r
# GEC estimator using EL(empirical likelihood) divergence
cal_EL <- GEcalib(~ 0 + Xs, dweight = d, const = total,
       method = "GEC0", entropy = "EL")
head(cal_EL$w)
```

```
##         1         2         3         4         5         6
## 54.743353 27.066422 17.977452 13.458167 10.754606  8.232996
```

```r
GECal::estimate(y ~ 1, calibration = cal_EL)$estimate
```

```
##   Estimate Std. Error
## y 1209.387   89.30535
```

1.1.3 Shifted KL(Cross entropy)

```r
# # GEC estimator using CE(cross entropy or shifted KL) divergence
cal_CE <- GEcalib(~ 0 + Xs, dweight = d, const = total,
       method = "GEC0", entropy = "CE", weight.scale = 2)
# design weights should be greater than 1 in CE
head(cal_CE$w)
```

```
##        1        2        3        4        5        6
## 54.88812 26.95156 17.91845 13.45263 10.78924  8.18484
```

					weights		
i	y_i	x_{i0}	x_{i1}	x_{i2}	**ET**	**EL**	**CE**
1	5	1	1	1	48.36	54.74	54.89
2	4	1	1	3	31.83	27.07	26.95
3	7	1	1	5	20.95	17.98	17.92
4	9	1	1	7	13.79	13.46	13.45
5	11	1	1	9	9.07	10.75	10.79
6	10	1	0	6	10.48	8.23	8.18
7	13	1	0	7	8.50	7.65	7.63
8	12	1	0	8	6.89	7.14	7.14
9	15	1	0	9	5.59	6.69	6.71
10	15	1	0	10	4.54	6.30	6.34

Table 2: Comparison of ET, EL, and CE(shifted KL) weights.

```
GECal::estimate(y ~ 1, calibration = cal_CE)$estimate
```

```
##   Estimate Std. Error
## y 1209.839   89.42995
```

2 Real data example

- The 2021 NHID from NHIS(National Health Insurance Service, Republic of Korea) was used, containing data for 1,000,000 adults in South Korea.
- A pseudo-population of size $N = 100{,}000$ was created via random sampling.
- Samples of size $n = 2{,}297$ were drawn using stratified sampling across 476 strata defined by Region (17), Age Group (14), and Sex (2).
- Stratum-specific sample sizes are:
  - $n_h = 5$ if $N_h > 15$, and $n_h = \lfloor N_h/3 \rfloor$ if $N_h \leq 15$
- We try to estimate the population sum of three study variables:
  - Hemoglobin level(`Hemo`, in g/dL), Oral examination status(`OralExam`, 1 if an oral examination was conducted, 0 otherwise), and Alcohol consumption status(`Alcohol` 1 or 0).
- See Kwon et al. [2024] for more details.

```
load("nhis.Rdata")

head(nhis.samp[,c("AgeGroup", "SEX", "REGION1", "Hemo", "Alcohol", "OralExam")])
```

```
##   AgeGroup SEX REGION1 Hemo Alcohol OralExam
## 1       16   2      27 12.5       0        0
## 2       16   2      27 11.2       0        0
## 3       16   2      27 11.4       0        0
## 4       16   2      27 14.1       0        1
## 5       16   2      27 13.2       0        0
## 6       16   1      27 13.7       0        1
```

```
fortmp <- formula(~ AgeGroup + SEX + REGION1)
const = colMeans(model.matrix(fortmp, nhis))
# const = colSums(model.matrix(fortmp, nhis)) # Used for population total

calibration <- GEcalib(
  fortmp,
  dweight = rep(1, nrow(nhis.samp)),
  data = nhis.samp,
  const = const,
  entropy = "ET",
  method = "GEC0"
)

estimate(Hemo + Alcohol + OralExam ~ 1, data = nhis.samp,
         calibration = calibration)$estimate
```

```
##             Estimate Std. Error
## Hemo      14.1310603 0.03121010
## Alcohol    0.6551861 0.01083104
## OralExam   0.3586279 0.01250435
```

References

Yonghyun Kwon, Jae Kwang Kim, and Yumou Qiu. Generalized entropy calibration for analyzing voluntary survey data. *arXiv preprint arXiv:2412.12405*, 2024.