11 praktikum – üldistatud regressioonihinnang

Tänases praktikumis vaatleme, kuidas kaasata registris leiduvat informatsiooni hinnangusse eesmärgiga parandada hinnangu täpsust. Selleks kasutame üldistatud lineaarset hinanngut (GREG), alternatiivset kalibreerimislähenemist on põhjalikumalt uuritud aines *Valikuuringud II*.

11.1 Lisainformatsiooni leidmine

Lisainformatsioon (*auxiliary information*) – teadaolevad kogusummad üldkogumi kohta (saadav väljaspoolt uuringut, näiteks mitmesugustest registitest).

StatVillage kohta teame

- $M = 1669$ inimeste arv;
- $N = 480$ leibkondade arv (seda tavaliselt ei teata);
- soo-vanusejaotus Rahvastikuregistris;
- ehitusaasta-omanike jaotus Eluasemeregistris;
- eluasemete jaotus geograafiliste piirkondade kaupa.

Lisainformatsiooni saamiseks registritest saame leida sagedustabelid. Näiteks,

```
proc freq data=popreg;
tables sugu*vanuskl/nopercent norow nocol;
run;
```

Sagedustabelist näeme, et Rahvastikuregistris on olnud mõned vead: on tekkinud sugu 0. Kes need on? Ehk alla 5 aastased lapsed. Liidame antud praktikumis need lastega:

```
data popreg1;
set popreg;
if sugu=0 then sugu=3;
run;
```

Agregeeritud lisainformatsioonina võib kasutada tabeli kõiki ruudusagedusi, või üksnes marginaalseid sagedusi, või sagedusi, mis tekivad klasside sobival kokkuvõtmisel.

Näide 1. Kasutades tabeli marginaalset soojaotust on lisainfoks $\boldsymbol{t}_x = \sum_U \boldsymbol{x}_i = (803, 746, 120)'$. Vektor $\boldsymbol{x}_i$ sisaldab i. leibkonnale vastavaid abitunnuste väärtuseid, ehk $\boldsymbol{x}_i$ = (meeste arv, naiste arv, laste arv)$'$. Kasutades vanuseklassi marginaalset jaotust saaksime veel ühe vektori $\boldsymbol{t}_x$. Lisainfot saame ka kombineerida omavahel.

Näide 2. Kasutades eluasemeregistrit saaksime lisainfot samuti perede kohta:

```
proc freq data=housereg;
tables builth*tenurh/nopercent norow nocol;
run;
```

Näeme, et väheste tunnuste baasil on võimalik moodustada pikk lisainformatsiooni vektor.

ÜLESANNE 1 Koosnegu üldkogum kõikidest StatVillage leibkondadest. Kirjutada välja vektori $\boldsymbol{t}_x$ väärtused, kus $\boldsymbol{x}_i =$ (meeste arv peres, naiste arv peres, leibkonna suurus, konstant 1).

11.2 GREG hinnang

Tuletame meelde, et üldistatud regressionihinnang TTA disainide korral on esitatav järgmiselt:

$$\hat{t}_{greg} = \hat{t}_y + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)'\hat{\boldsymbol{B}}, \tag{1}$$

kus

$$\hat{\boldsymbol{B}} = \Big[\sum_s \frac{\boldsymbol{x}_i \boldsymbol{x}_i{}'}{\sigma_i{}^2 \pi_i}\Big]^{-1} \sum_s \frac{\boldsymbol{x}_i y_i}{\sigma_i{}^2 \pi_i} = \hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{t}}_{xy}. \tag{2}$$

Seda saab kirja panna alternatiivselt uue kaalude süsteemi abil järgmiselt:

$$\hat{t}_{greg} = \sum_s w_i^r y_i, \tag{3}$$

kus

$$w_i^r = \frac{1}{\pi_i} g_{is} \quad \text{– regressioonkaal;}$$

$$g_{is} = 1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)'\hat{\boldsymbol{T}}^{-1}\frac{\boldsymbol{x}_i}{\sigma_i{}^2} \quad \text{– g-kaal.} \tag{4}$$

Oma arvutuskäikides kasutame valemit (4), mida on võimalik teisendada ka järgmisele kujule (eeldusel, et $\sigma_i = \sigma \; \forall i \in U$), mis annab korraga tervet g-kaalude vektorit:

$$g_s = \mathbf{1} + \boldsymbol{X}(\breve{\boldsymbol{X}}'\boldsymbol{X})^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x), \tag{5}$$

kus $\boldsymbol{X}$ on abitunnuste maatriks valimis ja $\breve{\boldsymbol{X}}$ on disainikaaludega läbikorrutatud abitunnuste maatriks valimis.

ÜLESANNE 2. Kasutame projektitöö I osas saadud valimit (LJV TTA). Lisame sinna tunnuseid, mis vastavad maatriksile $\boldsymbol{X}$ ülesandest 1:

```
data valim1;
set valim;
x1=hhperb1+hhperd1+...;\*meeste arv*\
x2=...;\*naiste arv*\
x3=hhsize;
x4=1;
kaal=480/100;\*juhul kui pole olemas*\
run;
```

Valime uuritavateks tunnusteks *nuirh* (sissetuleku saajate arv leibkonnas) ja *roomh* (tubade arv). Eeldame, et $\sigma_i = \sigma \; \forall i \in U$.

Järgmine protseduur võimaldab leida GREG hinnangud nendele tunnustele:

```
proc iml;
start;
t={803 746 1669 480}; *siia meie abiinfo kogusummad;
use valim;
read all var {nuirh roomh} into Y; *uuritavad tunnused;
read all var {kaal} into W; *kaalutunnus;
read all var {x1 x2 x3 x4} into X; *abitunnuste fail;
XL=W#X; *korrutab kaaluga kõik X veerud;
th=XL[+,]; *leiab veergude summasid, mis ongi hinnangud t_x-le;
g=1+X*inv(t(XL)*X)*t(t-th); *transponeerimine on t ja muutuja t;
Y_kaalutud=W#g#Y; *märk # elemendiviisilise korrutise jaoks;
greg=Y_kaalutud[+,]; *hinnangud mõlemale parameetrile korraga;
print greg;
finish;
run;
quit;
```

Selleks, et leida saadud hinnangutele dispersioonihinnanguid, peaksime oma iml protseduuri täiendama (jätame jätkukursuse jaoks). Siin aga vaatleme SAS-i olemasolevat protseduuri SURVEYREG, mis võimaldab üsna lihtsalt saada nii GREG hinnanguid kui ka vajalike täpsusenäitajaid. Dispersioonihinnangud leitakse lähendmeetoditel.

Järgmised kaks protseduuri on analoogsed (võrdle iml tulemusega!):

```
proc surveyreg data=valim;
model nuirh = x1 x2 x3;
weight kaal;
estimate 'sissetuleku saajate arv' x1 803 x2 746 x3 1669 intercept 480;
run;
```

või

```
proc surveyreg data=valim;
model nuirh = x1 x2 x3 x4/noint;
weight kaal;
estimate 'sissetuleku saajate arv' x1 803 x2 746 x3 1669 x4 480;
run;
```

ÜLESANNE 3. Võrdle kahe hinnangu dispersioonid LJV TTA dispersioonidega. Kas saab öelda, et GREG hinnangu korral on dispersioon väiksem?

ÜLESANNE 4. Vali oma projektitöö I osast veel 1-2 tunnust, mille hinnanguid sooviksid dispersiooni mõttes parandada GREGi abil. Mõtle, millised abitunnused võiksid olla sel juhul head. Kasutades protseduuri surveyreg leia nii hinnangud kui ka täpsusenäitajad. Võrdle projektitöö omadega. Kirjuta saadud tulemustest Moodle Wikisse (14. nädal).