Miből élünk? - Módszertan

Jankó Balázs KSH Módszertani főosztály

2018. november 28.

Miből élünk? - Mintavétel

- **Szempontok a tervezésnél:**

  - Reprezentatív, várhatóan minimális szórást biztosító minta
  - Kb. 6 ezer megvalósult háztartás (MNB)
  - 300-350 településen (LAF)
  - Településenként legalább 10 háztartás (LAF)
  - A vagyonos háztartásokat reprezentáljuk felül (MNB)
  - Budapest + Pest megye együttes aránya max. 45% (LAF).

Mintavétel

- Célsokaság: Magyarországon magánháztartásokban élő népesség
- Kiválasztási egység: lakás. A KSH címregiszterének nem foglalt címei közül (kivéve max 10 címes települések)
- Kétlépcsős rétegzett véletlen minta:
  - 1. lépcső: települések:
    - Önreprezentáló települések: Budapest kerületei, a legalább 10 ezer lakásos települések, a legalább 10 db. 80+ millió Ft becsült értékű lakást tartalmazó települések.
    - Rétegzett véletlen településminta nagysággal arányos kiválasztási valószínűséggel. Rétegek: régió (7) x SZJA-alap (4, kivéve Pest megye) x nagyságkategória (3). 81 rétegből 3-3 település PPS kiválasztása (összesen 243).

Mintavétel

- Összesen aránylag sok, 345 db település

- 2. lépcső: lakáscímek: a lakás értéke alapján 6 réteg. Réteghatárok: „kumulatív gyök f” szabály.
- Neyman vagy arányos allokáció -> „kompromisszumos” allokáció: még mindig hatékony, de elkerüli a budapesti meghiúsulást. Egyben megvalósítja a gazdagok felülreprezentálását.

Mintavétel - allokáció

A réteg sorszáma	A lakás becsült értéke (millió Ft)	A lakások száma országosan	6000 címes minta megoszlása
1	< 6	1370263	1454
2	6 - 12	1374074	1543
3	12 - 20	906380	1146
4	20 - 35	530074	929
5	35 - 80	206894	707
6	80 <	18007	221

Mintavétel

- A várható meghiúsulást is figyelembe kell venni: meghiúsulási szorzókat alkalmaztunk a legutóbbi felvételből (és egy-két feltételezésből) kiindulva: országrész x településtípus x lakásréteg bontásban, 1/5 –ös minimumot feltételezve.
- Ez mindig a tervezés legbizonytalanabb része, ezért pótminta bevetésére is felkészültünk. 15 ezer kijelölt címből végül – megfeszített terepmunkával – 5968 sikeres kérdőív jött be, így nem lett pótminta.
- A településrétegek alapesetben az egyes háztartásrétegek címeiből való sokaságbeli részesedésüknek megfelelően osztoztak a kijelölt címek között (kivétel: 10 címes korlát). A rétegek települései ugyanannyi címet kaptak.

Mintavétel - lakásrétegek

- A lakás szintű rétegzéshez rétegző változót kell rendelnünk a címregiszter összes címéhez
- Célszerű a lakás értékével próbálkozni: jól korrelál a vagyonnal és van esélyünk előállítani
- A NAV lakásforgalmi adataiból (2013-2016) indultunk ki: 429 ezer tranzakció
- Minden lakás utolsó árát 2016-ra vetítettük (átlagos megyei négyzetméterár-változás).
- A NAV állomány nem tartalmaz használható címazonosítót: több lépésben végzett címtisztítás után albetét, tömb és közterület szinten kapcsoltuk a címregiszterhez. Rendre 120 ezer (2,7%), 1,28 millió (29%), 3,83 millió (87%: közterületi szintű átlagár)

Mintavétel - lakásrétegek

- Népszámlálási és Tstar adatokat is a regiszterhez kapcsoltunk
- Ahol volt legalább építési terület szintű adat (1,28 millió), ott az a végleges becslés. Egyben ezen a halmazon becsültük a regressziókat.
- Mindenhol olyan stepwise lineáris regresszió, amihez volt adat: Tstar + Népsz. + NAV (közterületi): 2,54 millió, 582 ezer esetben vagy Népsz. vagy NAV. 2 ezer esetben (0.05%) csak Tstar.
- Korlátok: közterületi átlagár /3 vagy *3, ha nincs: 30-400e Ft.
- Rétegzéshez elfogadható pontosságot biztosít

Súlyozás

- Egy véletlen minta súlyozása már a mintavételnél kezdődik: Az ún. design súly a tartalmazási valószínűség reciproka: $w^d=1/\pi$.
- A minta erőssége itt kellemetlenséget okozott: a design súlyok (és a meghiúsulások) erősen szóródtak.
- A meghiúsulás miatt ez még nem becslősúly, ahhoz a design súlyokat korrigálni kell. A kerethibától megtisztított mintát felosztjuk a válaszadás szempontjából homogén csoportokra („cs”). A design súlyokkal válaszolási arányt számítunk: $\mathrm{val}_{cs} = \sum_{h \epsilon cs}(w_h^d * v_h) / \sum_{h \epsilon cs} w_h^d$. (v_h=1, ha a h-adik háztartás együttműködött).
- $w_h^1 = (\frac{1}{val_{cs}}) * w_h^d$

Súlyozás

- Probléma: a nem válaszolókról is tudnunk kellene, hogy melyik „cs” csoportba tartoznak. A keret rendszerint kevés információt tartalmaz: településréteg x lakásréteg szerinti csoportokat képeztünk.
- Egy-két változó mentén korrigálni még kevés lehet, ezért kalibrálással tovább igazítottuk a súlyokat (torzítás, szórás, célértékek szempontjából is hasznos)
- Feltételes optimalizálási feladat:

$$\sum_{k \in S} w_k x_k = x_U$$

$$E\left[\sum_{k \in S} G(w_k, d_k)\right] \rightarrow min$$

Súlyozás

- GREG becslés: $G(w,d) = \frac{1}{2}\frac{(w-d)^2}{d} = d\frac{1}{2}\left(\frac{w}{d}-1\right)^2$
- Raking ratio: $G(w,d) = w\,log\left(\frac{w}{d}\right) + d - w = d\left[\frac{w}{d}log\left(\frac{w}{d}\right) + 1 - \frac{w}{d}\right]$.
- *Deville, Jean-Claude – Särndal, Carl-Erik (1992): Calibration Estimators in Survey Sampling.* http://www.jstor.org/stable/2290268

- Sarokszámok (X_U): pontosnak gondolt forrásokból: népszámlálás-továbbvezetés, MEF(?).:
- demográfiai változók 7 régió szerint: 0-17 éves férfiak és nők, 18-34 éves férfiak/nők, 35-54 éves férfiak/nők, 55-74 éves férfiak/nők, 75 éves vagy idősebb férfiak/nők száma. (Összesen: 5*2*7=70 db.)

Súlyozás

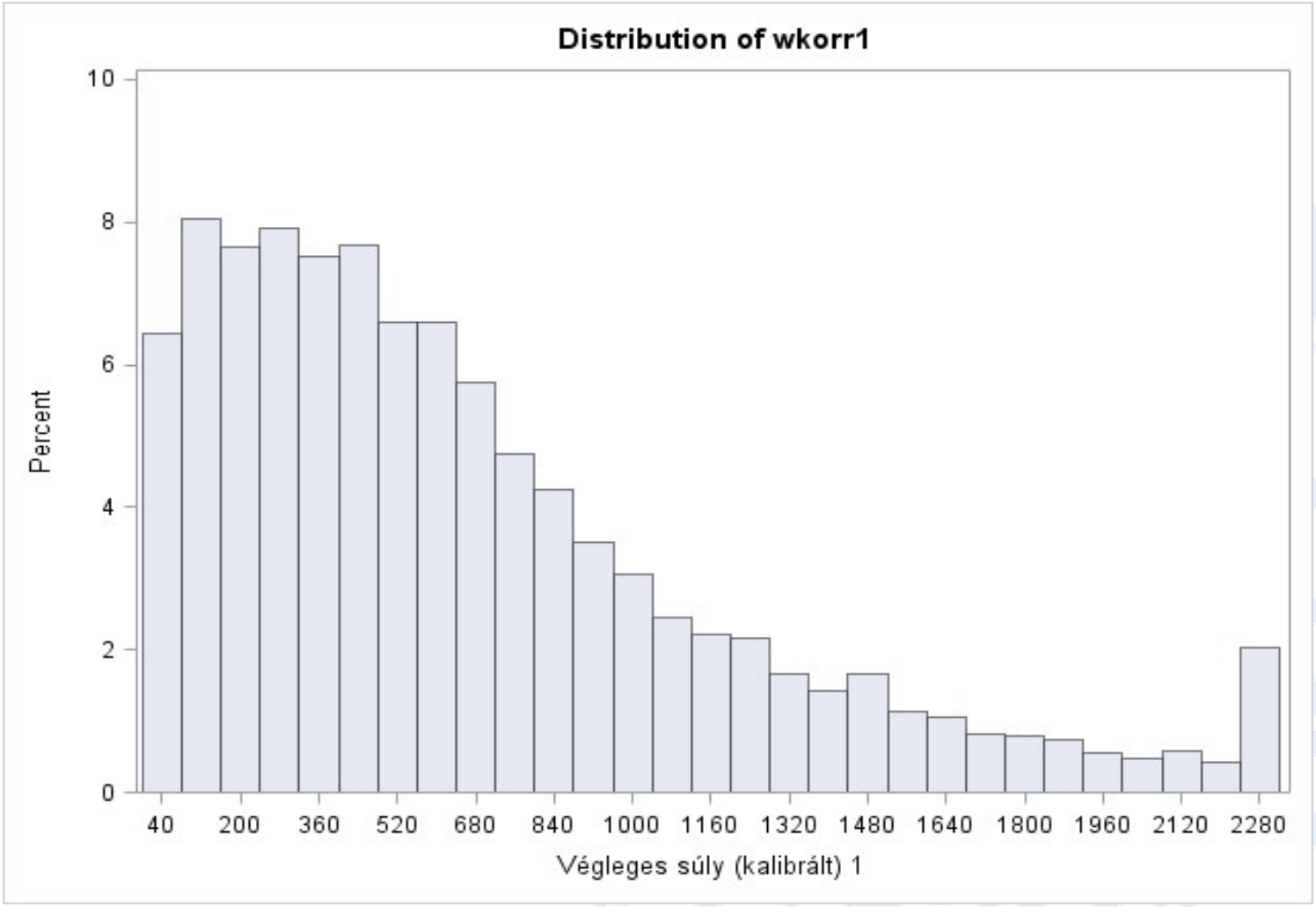
- Aktivitási, végzettségi változók régiónként: alkalmazottak, vállalkozók, munkanélküliek, nyugdíjasok, egyéb inaktívak, legfeljebb alapfokú végzettségűek, érettségi nélküli középfokú végzettségűek, érettségizett középfokú végzettségűek, felsőfokú végzettségűek, Budapestiek, megyei jogú városban lakók, kisebb városban lakók, falusiak. Összesen: 13*7=91 db. (ezek vannak a kérdőíven és van róluk sarokszám).
- Felváltva kalibráltunk a halmazokhoz, 50 / 2300 –as korlátot alkalmazva: kicsit széles intervallum.
- Cím szintű súlyok

KÖZPONTI STATISZTIKAI HIVATAL

Súlyozás - illeszkedés

Sarokszám	KOZP becslés/ cél	KDU becslés / cél	NYDU becslés / cél	DDU becslés / cél	EMO becslés / cél	EAL becslés / cél	DAL becslés / cél
ffi 0-17	100.10%	100.00%	99.30%	100.80%	100.20%	100.40%	99.76%
ffi 18-34	99.92%	99.87%	99.89%	99.91%	99.84%	99.78%	99.88%
ffi 35-54	99.92%	99.71%	100.30%	99.73%	99.91%	99.65%	100.00%
ffi 55-74	99.96%	99.92%	99.93%	100.00%	100.00%	99.87%	99.98%
ffi 75-x	99.94%	99.94%	99.76%	100.10%	99.98%	100.10%	99.91%
nő 0-17	99.97%	100.10%	99.24%	100.60%	100.10%	100.40%	99.62%
nő 18-34	100.00%	99.93%	100.30%	99.66%	99.89%	99.87%	100.20%
nő 35-54	99.99%	100.20%	100.40%	99.65%	99.99%	100.00%	100.20%
nő 55-74	100.10%	100.20%	100.10%	99.93%	100.10%	100.00%	100.10%
nő 75-x	100.00%	100.10%	99.93%	100.10%	100.00%	100.10%	99.95%

Súlyozás - súlyeloszlás



Megvalósulás

Réteg	Megvalósulás
1	0.63
2	0.49
3	0.42
4	0.36
5	0.35
6	0.32
Településtípus	**Megvalósulás**
Budapest	0.31
Megyei j. város	0.43
Egyéb város	0.51
község	0.6

Megye	Megvalósulás
Budapest	0.31
Borsod-A-Z	0.64
Csongrád	0.38
Győr-M-S	0.41
Heves	0.73
Pest	0.39
Szabolcs-Sz-B	0.64
Tolna	0.69

Megvalósulás

Településtípus	Réteg	Megvalósulás
Budapest	1	0.29
	2	0.27
	3	0.31
	4	0.3
	5	0.32
	6	0.32
Megyei j. város	1	0.37
	2	0.48
	3	0.45
	4	0.41
	5	0.41
	6	0.3

Településtípus	Réteg	Megvalósulás
Egyéb város	1	0.58
	2	0.54
	3	0.48
	4	0.43
	5	0.44
	6	0.38
község	1	0.72
	2	0.58
	3	0.53
	4	0.43
	5	0.33
	6	0.25

Megvalósulás

Megye	Réteg	Megvalósulás
Pest	1	0.46
	2	0.41
	3	0.37
	4	0.39
	5	0.36
	6	0.34
Heves	1	0.86
	2	0.74
	3	0.61
	4	0.66
	5	0.63
	6	0.5
Borsod-A-Z	1	0.72
	2	0.57
	3	0.44
	4	0.42
	5	0.75
	6	0.2

KÖZPONTI STATISZTIKAI HIVATAL

Imputálás

- Item nonresponse -> Multiple imputation módszer:
- MAR esetén a(z egyik) legjobb
- Az irodalomban népszerű, a gyakorlatban kevésbé (bonyolult, több adatbázist állít elő)
- Y: imputálandó változó(k), X: segédváltozó(k)
    - megbecsüljük az X és Y átlagait és kovariancia-mátrixát
    - Ezekből (lineáris esetben) adódnak a Y = f(X) = a + bX + e sztochasztikus regressziók
    - f(X)-el (X nem hiányzó részhalmazával) imputáljuk Y-t

Imputálás

- az átlagoknak(**μ)** és a kovariancia mátrixnak (**Σ)** is előállítjuk az eloszlásait: poszterior eloszlások=skálatényező*prior*likelihood
- Normális eloszlásúnak feltételezett Y és független mintavétel mellett az átlagok (többdimenziós) normális, a variancia-kovariancia mátrix elemei inverz Wishart eloszlást követnek.
- A fenti eloszlásokból generálunk **Σ** és **μ** vektorokat, amik segítségével kiszámíthatók az új sztochasztikus regressziók paraméterei
- Sok ilyen iterációt végzünk

Imputálás

- Különféle eszközök (pl. autokorreláció-függvény) segítségével meghatározzuk azokat a becsléseket, melyek már elég távol vannak egymástól ahhoz, hogy függetlennek tekinthessük
- MI pontbecslés: az iterációk átlaga
- MI variancia-becslés: szokásos módon számolt varianciából és pontbecslésekből adódik:
- Bootstrap-variancia: $U(\hat{\theta})_m = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_{mb} - \overline{\theta_m})^2, B = 1000, M = 5.$
- *Teljes variancia:* $\frac{1}{M}\sum_{m=1}^{M} U(\hat{\theta})_m + \left(1 + \frac{1}{M}\right)\sum_{m=1}^{M}(\hat{\theta}_m - \overline{\theta})^2$