Adatcserével anonimizált mikroadatok használhatósága – Egy szimulációs vizsgálat tanulságai*

Bartus Tamás
PhD, a Budapesti Corvinus Egyetem docense
E-mail: tamas.bartus@uni-corvinus.hu

A tanulmány áttekinti az adatok felfedés elleni védelmét szolgáló eljárások statisztikai következményeit, és részletesen elemzi az adatcsere kovariancia- és regressziós becslésekre gyakorolt hatását. Amellett érvel, hogy az adatcsere kitüntetett szerepet tölt be a felfedés elleni védelem során. Az adatcsere különböző módszereinek kovariancia- és regressziós becslésekre gyakorolt hatását szimulációval vizsgálja. E vizsgálatok eredménye szerint az esetek többségében az adatcseréből fakadó relatív torzítás mértéke 10 százalék alatt tartható. A torzítást egyrészt a donorok véletlenszerű kiválasztása, másrészt az adatcsere (magyarázó-) változók közötti megosztása minimalizálja. Az eredményeket a mérési hibák elméletére támaszkodva értelmezi.

TÁRGYSZÓ:
Mikroadatok.
Adatcsere.
Anonimizálás.

* A tanulmány az Új Magyarország Fejlesztési Terv Társadalmi Megújulás Operatív Program támogatási rendszeréhez benyújtott „Munkaerő-piaci előrejelzések készítése, szerkezetváltási folyamatok előrejelzése” című TÁMOP-2.3.2-09/1-2009-0001 kiemelt projekt keretében készült. Köszönettel tartozom *Daróczi Gergőnek* lelkiismeretes asszisztensi munkájáért. Szintén köszönet illeti *Cseres-Gergely Zsombort* a kutatást segítő megjegyzéseiért, valamint azért, hogy rendelkezésemre bocsátotta a KSH munkaerő-felvétel 2011. I. negyedéves anonimizált adatait. A tanulmány korábbi változatát „A mikroadatok hozzáférésével és az adatok felfedés elleni védelmével kapcsolatos kérdésekről” (Budapest, 2012. november 6.) szervezett műhelykonferencián prezentáltam. Hasznos észrevételeikért köszönettel tartozom a KSH Módszertani főosztálya munkatársainak: elsősorban *Szép Katalinnak* és *Vereczkei Zoltánnak*, valamint *Antal Lászlónak*, *Dobány Máténak* és *Nagy Beátának*.

A statisztikai törvény és annak végrehajtási rendelete az adatszolgáltató beleegyezéséhez kötötte az ún. egyedi, azaz az adatszolgáltatóval „kapcsolatba hozható" adatok továbbadását.[1] A jogszabályok miatt az adatgazdák – például a Központi Statisztikai Hivatal (KSH) – az adatokat csak anonimizálás után adhatják tovább. Az adatszolgáltatók védelmére számos eljárást dolgoztak ki (*Hundepool et al.* [2010]); ezekről magyarul is tájékozódhatnak az érdeklődők (*Bánszegi* [1997], *Erdei–Horváth* [2004], *Szép–Gadácsi* [2010]). A hatásos adatvédelem azonban korlátozza a felhasználók érdekeit (*Boudreau* [2005]), sőt akár ellehetetlenítheti a fontos kérdések empirikus kutatását. Kutatói szempontból nem az adatvédelem hatásossága, hanem az anonimizált adatok használhatósága a fő kérdés; az, hogy milyen mértékben veszélyezteti a hatásos adatvédelem az adatokból levont következtetések érvényességét. Ezzel a kérdéssel a hazai szakirodalom mindeddig alig foglalkozott. Tanulmányunk ezért az anonimizálásból fakadó torzítással foglalkozik.

A célunk ennél konkrétabb: a felfedés elleni védelem egyik eljárásának, az adatcserének a kovariancia- és regressziós becslésekre gyakorolt torzító hatását vizsgáljuk. Az adatcserére több okból esett a választás. Egyrészt az anomimizálási eljárások vagy technikák célja az, hogy ne lehessen az adatszolgáltatóval „kapcsolatba hozni" az adatbázisokban található információkat. Az adatszolgáltatókat ún. kváziazonosítók segítségével lehet azonosítani: ezek olyan könnyen megfigyelt változók (például lakóhely, nem, korcsoport, gyerekszám), melyek együttesen alkalmasak a megfigyelt egyén felfedésére. Az adatcsere pont a kategorikus változók védelmére szolgál. Másrészt – ahogy nemsokára látni fogjuk – a felhasználók szemszögéből az adatcsere számos kedvező tulajdonsággal rendelkezik. A többváltozós regresszió-elemzés során bekövetkező torzításokat azért érdemes vizsgálni, mert az alapkutatások és a hatásvizsgálatok talán legfontosabb adatelemzési módszere. Habár az adatvédelmi technikákat áttekintő publikációkban (*Domingo-Ferrer–Torra* [2001a], [2001b]; *Hundepool et al.* [2010]) számos, az információveszteséget mérő általános mérőszámot találhatunk, a regressziós együtthatók torzulásával kapcsolatos konkrét eredmények hiányoznak.[2] A szakirodalom tárgyalta egyes technikák variancia- és kovarianciabecslésekre gyakorolt hatását, sőt azt is, hogyan lehet anonimizált állományokból torzítatlan variancia- és kovarianciabecsléseket végezni (*Kim* [1990], *Gouweleeuw et al*. [1998]). Kérdéses azonban, hogy az eredmények kiterjeszthetők a többváltozós becslések kontextusára.

[1] Lásd az 1993. évi XLV. I. törvény 17. paragrafusát, valamint a 170/1993. (XII. 3.) Korm. rendelet 16. paragrafusát. A hazai jogszabályok célja egybeesik például az amerikai gyakorlattal; lásd *Sullivan* [1992].

[2] A regresszióelemzés iránti érdektelenség valószínűleg azzal magyarázható, hogy az adatvédelem a hivatalos statisztika része, a statisztikai hivatalok munkatársainak feladata pedig nem a regresszióelemzés, hanem átlagok és szórások publikálása, illetve becslése.

A tanulmány újdonsága a szimulációs módszer használata. A többváltozós modellek legkisebb négyzeteken alapuló becslése ugyanis a magyarázóváltozók variancia-, kovarianciamátrixa inverzének és a függő és a magyarázó változók kovarianciamátrixának (pontosabban vektorának) szorzata. A mátrixalgebra miatt nehezen látható át, milyen mértékben torzulnak a becslések, ha az anonimizálás miatt módosul egyes magyarázóváltozók varianciája vagy kovarianciája. A probléma hasonló ahhoz, amikor mérési hiba folytán egy adott változó szórása nő, és ezáltal az összes változó együtthatójának regressziós becslése módosul (*Maddala* [2004]). Ez a nehézség indokolja a szimulációs módszer használatát. A szimulációs módszert eddig a mikroaggregálás regressziós becslésekre gyakorolt hatásainak elemzésére használták (*Liu–Little* [2003]; *Lenz et al.* [2006]; *Schmid–Schneeweiss* [2005], [2007], [2008]).

A tanulmány felépítése a következő. Először áttekintjük a mikroadatok anonimizálására széleskörűen használt eljárásokat és ezek átlag-, szórás- és kovarianciabecslésekre gyakorolt hatásait. Mivel az adatvédelmi szabályok a kváziazonosítók anonimizálására ösztönöznek, a kváziazonosítók – mint például a településkód, a foglalkozás kódja – pedig gyakran kategorikus változók, az adatvédelem egyik legfontosabb technikájának az adatcserének kell lennie. A második rész az adatcsere technikáit, illetve az adatcserével kapcsolatos eredményeket elemzi. Bemutatjuk, hogy az adatcsere kovarianciabecslésekre gyakorolt hatása a mérési hibák fogalmi keretén belül értelmezhető. Ez az eredmény azért fontos, mert a mérési hibák többváltozós regressziós becslésekre gyakorolt hatása analitikusan nehezen kezelhető (*Maddala* [2004]), az esetleges torzítások vizsgálata ezért szimulációra váró téma. A harmadik rész a szimulációs vizsgálat módszerét és az eredményeket értelmezi. A szimulációhoz a KSH munkaerő-felvételének 2011. első negyedéves (anonimizált) adatait használjuk. A regressziós becslésekkel kapcsolatos vizsgálatok során egy olyan kutató helyzetét vizsgáljuk, aki a lakóhely-azonosítót is tartalmazó adatbázist szeretné használni, de adatvédelmi okok miatt az adatgazda csak akkor bocsátja ezt rendelkezésre, ha az azonosításhoz szükséges egyéb változókat módosítják. A szimuláció során különböző feltevéseket fogalmazunk meg azzal kapcsolatban, hogy kik azok, akik védelemre szorulnak, és milyen változók módosításával garantálható az anonimitás. A tanulmány végén az eredményeket a mintavételi és a mérési hibák elméletére támaszkodva értelmezzük.

1. Mikroadatok anonimizálásának statisztikai következményei

A felfedés elleni védelem – különösen a hazai jogszabályok fényében – legegyszerűbb módja a kváziazonosítók visszatartása (törlése) vagy átkódolása. Ezek a

technikák nem torzítják, hanem lehetetlenné tesznek bizonyos becsléseket, továbbá jelentősen korlátozzák az adatbázis használhatóságát. Ebben a szakaszban áttekintjük az ennél kevésbé korlátozó, de az adatok módosításával járó technikák statisztikai következményeit. Konkrétan azt vizsgáljuk, befolyásolják-e az egyes eljárások az anonimizált – azaz valamilyen adatvédelmi technikával módosított – változók átlagát, szórását és más változókkal számolt kovarianciáját.[3] Ezek a statisztikák alkotják az adatelemzés során leggyakrabban használt eljárások – például a regresszióelemzés, a faktorelemzés – inputjait. Az eljárások logikájával, alapjaival foglalkozunk, és figyelmen kívül hagyjuk az egyes eljárásokon belüli további technikai változatokat, melyek célja az adatvédelem hatásosságának fokozása. A fejezetben található képletek azt feltételezik, hogy a felfedés elleni technikákat a teljes adatbázison, nem pedig annak valamelyik részmintáján használják.

1.1. Adathiány-generálás

Az eljárás során az egyik kváziazonosító változó értékét a magas felfedési kockázatú egyéneknél adathiányra kódoljuk át, úgy, hogy a többi kváziazonosító már ne tegye lehetővé az azonosítást. A módszer kifinomultabb változata annak a „brutális" megoldásnak, amikor az egész esetet törlik az adatbázisból. A módszer nyilvánvaló hátránya az elemzéshez használható mintanagyság csökkentése és az anonimizált változó átlagának torzulása. Ha az n elemű mintában $k = pn$ megfigyelésnél töröljük az x változó értékét, akkor az anonimizált x^a változó átlaga

$$\bar{x}^a = \frac{\bar{x} - p\bar{x}_k}{1 - p}$$

lesz, ahol $\bar{x}_k$ x átlaga a törléssel védett részmintában, p pedig az anonimizált megfigyelések relatív gyakorisága. A képlet súlyozott átlagbecslésekre is érvényes. Ha a súlyok normalizáltak, azaz a súlyok összege azonos a mintanagysággal, akkor az egyenletben a p paramétert a törölt megfigyelésekhez tartozó súlyok összegének és a mintanagyság hányadosaként kell értelmezni – azaz:

$$p = n^{-1}\left(\sum_{(k)} w\right),$$

[3] A tanulmányban rendszeresen használjuk a változó anonimizálása, valamint az anonimizált változó terminusokat. Az előbbi a „változó adatvédelmi okok miatt végzett módosítása", az utóbbi az „adatvédelmi megfontolások miatt módosított változó" kifejezést rövidíti.

ahol w_i az i esethez tartozó súly, $\sum w$ pedig a súlyok összege. A képlet világosan mutatja, hogy az átlagbecslés akkor torzul, ha kis mintából kiugró értéket törlünk.

A $0-1$ kódolású indikátorváltozóknál még egyszerűbb a képlet. Ha törlésre csak az $x=1$ értékeknél kerül sor, $\bar{x}_k = 1$ és az anonimizált indikátorváltozó súlyozatlan átlaga

$$\bar{x}^a = \frac{\bar{x} - p}{1 - p}.$$

Az anonimizálás okozta torzítás

$$\bar{x}^a - \bar{x} = \frac{p}{1-p}(\bar{x} - 1),$$

tehát annál nagyobb, minél nagyobb p és minél nagyobb a változó átlaga. Az anonimizált indikátorváltozó szórásnégyzete:

$$Var\left(x^a\right) = \frac{\bar{x} - p}{(1-p)^2 \bar{x}} Var(x).$$

Ha p értéke nulla, az anonimizált változó és annak varianciája azonos az eredetivel. Mivel a variancia sosem lehet negatív, $p \le \bar{x}$. Mivel x indikátorváltozó, az egyenlőtlenség azt a triviális feltételt fogalmazza meg, hogy az anonimizált esetek aránya nem haladhatja meg az $x=1$ esetek arányát. A p növekedésével tehát az anonimizált változó varianciája csökken; az anonimizálás „elkoptatja" az eredeti változó varianciáját.

Szintén torzulhat az anonimizált indikátorváltozó egy tetszőleges másik változóval vett kovarianciája. Ha az adatvédelem az adatbázis 100 p százalékára terjed ki, és ismét csak $x=1$ esetekre, akkor az anonimizált indikátorváltozó és a tetszőleges anonimizálatlan y változó kovarianciája:

$$Cov\left(x^a, y\right) = \frac{Cov(x,y)}{(1-p)^2} - \frac{p(1-\bar{x})(\bar{y}_1 - \bar{y}_0)}{(1-p)^2},$$

ahol $\bar{y}_1$ és $\bar{y}_0$ y átlaga az anonimizálatlan adatbázisban az $x=1,$ illetve $x=0$ csoportokban. A kovarianciabecslés torzulása nyilvánvalóan p és a szóban forgó csoportátlagok közti különbség függvénye.

Az adathiány-generálásnál tehát egyszerű képletet kaptunk arra, milyen mértékben torzulnak az átlag- és varianciabecslések. A kovarianciabecslések torzulására kapott képlet viszont bonyolultabb.

1.2. Adatcsere

Az adatcsere (data swapping) során a felfedési kockázatot jelentősen növelő változó (vagy változók) értékeit cseréljük fel egyes megfigyelések között (*Dalenius–Reiss* [1982]). Képzeljük el, hogy egy adatbázisban magas a falusi egészségügyi dolgozók és a városi mezőgazdasági dolgozók felfedési kockázata. A felfedési kockázat csökkenthető, ha *k* számú falusi egészségügyi dolgozó foglalkozását mezőgazdasági dolgozóra, és ezzel párhuzamosan szintén *k* számú városi mezőgazdasági dolgozó foglalkozását egészségügyi dolgozóra módosítjuk – azaz a foglalkozási adatokat kicseréljük.

Legyen $\delta_{xij} = 1,$ ha az *x* változó értékét az *i* és a *j*-edik megfigyelések között kicseréljük; különben $\delta_{xij} = 0.$ Az adatcsere formális definíciója a következő (*Boudreau* [2005]):

$$x_i^a = \left(1 - \delta_{xij}\right) x_i + \delta_{xij} x_j ,$$

$$x_j^a = \left(1 - \delta_{xij}\right) x_j + \delta_{xij} x_i .$$

A formális definíció – meglepő módon – semmilyen információt nem tartalmaz az *i* és *j* egyének felfedési kockázatáról. Az adatcsere céljának figyelembe vétele mellett triviális, hogy a két egyén közül az egyik – de csak az egyik – könnyen felfedhető.

Az adatcsere nem módosítja az átlagot és a szórást, de nem őrzi meg feltétlenül az együttes eloszlásokat. Példánkban a foglalkozás cseréje után mind a településtípus, mind pedig a foglalkozás peremeloszlása változatlan marad – ugyanakkor megváltozik a foglalkozás és településtípus együttes eloszlása, hiszen a csere révén csökken a falusi egészségügyiek és a városi mezőgazdaságiak száma (és értelemszerűen nő a falusi mezőgazdasági és a városi egészségügyi dolgozók száma). A probléma megoldására *Dalenius* és *Reiss* [1982] azt javasolta, hogy az adatcserét további megfigyelések bevonásával kell folytatni, mindaddig, amíg helyreáll a többdimenziós eloszlás. A sikerre azonban nincs garancia; ráadásul az újabb és újabb cserék megtalálása roppant időigényes. A gyakorlatban is könnyen megvalósítható adatcsere ezért csak a peremeloszlásokat őrzi meg tökéletesen – az együttes eloszlásokat viszont csak közelítőleg (*Reiss* [1984]). Ilyen könnyen kivitelezhető technika például az, amikor az

adatcserébe bevont esetek a kicserélt változóktól eltérő más változók szempontjából hasonlítanak egymásra (*Shlomo–Tudor–Groom* [2010]).

Az együttes eloszlás változásának két következménye van. Egyrészt torzulnak a súlyozott becslések, hiszen az adatcsere nem terjed ki a súlyváltozókra. Ha az x_j és x_k értékeket cseréljük ki, akkor az anonimizált és az eredeti változók súlyozott átlagainak különbsége

$$\frac{w_j\left(x_k - x_j\right) + w_k\left(x_j - x_k\right)}{\sum w_i} \qquad /1/$$

lesz, ahol w_i az i-edik megfigyeléshez rendelt súly. A súlyozott becslésekre vonatkozó képletek rendkívül bonyolultak (*Boudreau* [2005]).

A másik következmény: torzulnak a (súlyozatlan) kovarianciák. Tegyük fel, hogy az adatcsere *pn* megfigyelés*pár*t érint. Az x változón végrehajtott adatcsere következményeit tekintve azonos az y változón végzett adatcserével. Jelölje y_{01} és y_{10} azokat az y értékeket, melyeknél az x indikátorváltozót nulláról egyesre, illetve egyesről nullára cserélték. Az adatcsere egyetlen hatása: x^a és y szorzatösszegét $\sum y_{01}$ összeggel növeljük és a $\sum y_{10}$ összeggel csökkentjük. Ez alapján az adatcseréből fakadó torzítás

$$Cov\left(x^a, y\right) - Cov\left(x, y\right) = -p\left[\bar{y}_{10} - \bar{y}_{01}\right]. \qquad /2/$$

Ha y várható értéke az $x = 1$ csoportban magasabb, és az anonimizált megfigyelések a minta véletlenszerűen kiválasztott mintája, akkor $\bar{y}_{10} > \bar{y}_{01}$ és így a /2/ egyenlet jobb oldalán szereplő különbség negatív. Ennek ellentéte igaz, ha y várható értéke az $x = 0$ csoportban magasabb. Az adatcsere tehát a csökkenti, „koptatja" a kovariancia abszolút értékét. A kopás mértéke annál nagyobb, minél több megfigyelésre terjed ki az adatcsere.

Mivel tetszőleges x indikátorváltozó és tetszőleges y változó kovarianciája x varianciájának és a $\bar{y}_1 - \bar{y}_0$ különbség szorzata, az anonimizált x^a indikátorváltozó és y kovarianciája a következő formára hozható:

$$Cov\left(x^a, y\right) = Cov\left(x, y\right)\left[1 - \frac{p}{Var\left(x\right)}\frac{\bar{y}_{10} - \bar{y}_{01}}{\bar{y}_1 - \bar{y}_0}\right].$$

A jobb oldalon a szögletes zárójelben szereplő mennyiséget érdemes külön jelöléssel ellátni:

$$Q_x(y) = 1 - \frac{p}{Var(x)} \frac{\bar{y}_{10} - \bar{y}_{01}}{\bar{y}_1 - \bar{y}_0}. \qquad /3/$$

Ha az adatcsere véletlenszerű és teljesül az $\bar{y}_{10} - \bar{y}_{01} = \bar{y}_1 - \bar{y}_0$ egyenlőség, $Q_x(y)$ még egyszerűbben írható fel:

$$Q_x(y) = 1 - \frac{p}{Var(x)}. \qquad /4/$$

A képlet üzenete világos: az adatcserével védett megfigyelések növekedésével egyre nagyobb mértékben torzul a kovariancia. A torzulás mértéke azonban az adatcserével érintett változó varianciájától is függ. Ha az adatgazda nyilvánosságra hozza a p együttható értékét, a felhasználó a

$$\widehat{Cov(x,y)} = \frac{Cov(x^a, y)}{Q_x(y)} \qquad /5/$$

képlettel becsülheti az anonimizálatlan állományban érvényes kovarianciát.

1.3. Utólagos randomizálás

Az *utólagos randomizálás* (post-randomization – PRAM) az adatcsere kifinomultabb változata: ez eljárás során adott változó értékeit egy előre meghatározott eloszlás szerint véletlenszerűen módosítják (*Kooiman–Willenborg–Gouweleeuw* [1997], *Gouweleeuw et al*. 1998). A módszert a randomizált válaszok technikája (*Sarndal–Swensson–Wretman* [1982]) inspirálta. Diszkrét, 0 és 1 értékeket felvevő változó esetén a módszer azt írja elő, hogy a nullákat adott $(1-\theta_0)$ valószínűséggel 1-re, az egyeseket adott $(1-\theta_1)$ valószínűséggel nullákra cseréljük. Többértékű változókra általánosítva: az adatcserét irányító valószínűség-eloszlást egy $k \times k$ dimenziójú **P** (perturbációs) mátrix definiálja, melynek *ij*-edik eleme annak valószínűségét adja meg, hogy a változó *i*-edik értéke kicserélődik a *j*-edik értékre.

Az adatok cseréje tehát nem a védelemre szoruló egyént, hanem a véletlenszerűen kiválasztott egyéneket érinti. Ha ezt a tényt az adatgazda nyilvánosságra hozza, a rosszindulatú felhasználó nem lehet biztos abban, hogy egy adott falu állatorvosa tényleg falusi állatorvos – hiszen lehetséges, hogy az utólagos randomizálás pont egy falusi mezőgazdasági segédmunkás foglalkozását cserélte fel egy városi állatorvos foglalkozására.

Az utólagos randomizálás után módosul a manipulált változó átlaga és szórása. Az utólag randomizált kétértékű változó átlaga

$$\bar{x}^a = (\theta_0 + \theta_1 - 1)\bar{x} + (1 - \theta_0), \qquad /6/$$

szórásnégyzete pedig

$$Var(x^a) = (\theta_0 + \theta_1 - 1)^2 Var(x) \qquad /7/$$

lesz (*Gouweleeuw et al*. [1998]). Az utólagos randomizálás tehát torzíthatja az átlagot és a varianciát. Az átlagok és a szórások azonban anonimizált állományokból is becsülhetők maradnak – feltéve, hogy az adatgazda nyilvánosságra hozza a randomizálás során használt **P** perturbációs mátrixot. Ha a **θ** paraméterek ismertek a felhasználók számára, a /6/–/7/ egyenletek alapján a korrigált átlagbecslés

$$\hat{\bar{x}} = \frac{\bar{x}^a - (1 - \theta_0)}{(\theta_0 + \theta_1 - 1)}, \qquad /8/$$

a korrigált varianciabecslés pedig

$$\widehat{Var(x)} = \frac{Var(x^a)}{(\theta_0 + \theta_1 - 1)^2} \qquad /9/$$

(*Gouweleeuw et al*. [1998]). A becslés természetesen csak akkor lehetséges, ha θ_0 és θ_1 összege nem azonos eggyel. Ha például úgy döntünk, hogy a diszkrét változó zérusait 5 százalékos eséllyel cseréljük egyre, akkor ezzel párhuzamosan nem dönthetünk úgy, hogy az egyes értékeket 95 százalékos eséllyel cseréjük nullákra.

A felhasználók érdekei akkor sérülnek legkevésbé, ha az utólagos randomizálás az esetszámra azonos (*Bycroft–Merrett* [2005] 126. old.). A **P** perturbációs mátrix elemeit ekkor úgy választjuk ki, hogy az anonimizált és az eredeti változó átlagai azonosak legyenek. Például kétértékű változók esetén akkor esetszámra azonos az utólagos randomizálás, ha teljesül az alábbi egyenlőség:

$$\theta_1 = 1 + \frac{1 - \bar{x}}{\bar{x}}(\theta_0 - 1).$$

Indikátorváltozóknál az esetszámra azonos utólagos randomizálás nem torzítja az átlagok és a szórások becslését, viszont a /2/ egyenlet miatt torzíthatja a kovarianciabecsléseket, illetve a súlyozott átlagbecsléseket.

1.4. Mikroaggregálás

A mikroaggregálásnak számos technikai változata létezik (*Mateo-Sanz–Domingo-Ferrer* [1998], *Schmid–Schneeweiss* [2005]). Az eljárás logikája mégis egyszerű. Első lépésben az adatokat a védelemre kiszemelt változó vagy egy másik változó szerint sorba rendezzük. Ezután a megfigyeléseket előre rögzített *k* vagy az eljárás során – valamilyen statisztikai eljárással megállapított – változó nagyságú csoportokba soroljuk. Az egyéni megfigyeléseket végül a szóban forgó csoportátlagokkal helyettesítjük, amelyeket a rendezés miatt egymáshoz hasonló adatokból számolunk ki, az anominizált és a valós értékek eltérése kicsi is lehet. Ez felveti azt a kérdést, vajon hatásosan védi-e a mikroaggregálás a személyes adatokat. Azonban a sorbarendezés nélküli csoportképzés sem garantálja automatikusan a hatásos védelmet: előfordulhat, hogy egy csoportspecifikus értékösszeget egyetlen megfigyelés dominál.[4]

Az eljárás nyilvánvalóan változatlanul hagyja a változó átlagát és csökkenti a varianciát: a szórásnégyzet-felbontás közismert képlete alapján az anonimizált változó varianciája a belső szórásnégyzettel, azaz a kategóriákon belüli szórásnégyzetek összegével lesz kisebb az eredeti változó varianciájánál. Az aggregálást megelőző sorbarendezés célja az, hogy a szórás csökkenése minimális legyen. A szórásbecslések mellett az eljárás a kovarianciabecsléseket is torzíthatja. Ha az eljárás során *j* darab *k* elemű aggregátumot alakítanak ki, akkor a torzítás – azaz az anonimizált és az anonimizálatlan állományokon számolt kovarianciák különbsége –

$$Cov\left(x^a, y\right) - Cov\left(x, y\right) = -\sum_{(j)} \sum_{(i=1)}^{K} y_{ki} \left(x_{ki} - \overline{x}_k\right).$$

Ha az egyes mikroaggregátumokon belül azonos lenne a megfigyelések csoportátlagoktól való eltérése, a torzulás mértéke a mikroaggregátumok méretének növekvő függvénye. A kovarianciák és a varianciák módosulása miatt a regressziós becslések is torzulnak – a torzítás konkrét mértékét számos szimulációs vizsgálatban elemezték (*Liu–Little* [2003]; *Lenz et al.* [2006]; *Schmid–Schneeweiss* [2005], [2007], [2008]).

1.5. Zajosítás

Az eljárás lényege: az egyedi vagy ritka adatokhoz egy véletlen zajt – azaz 0 átlagú, előre meghatározott szórással rendelkező ε véletlen számot – adunk. A zajosítás

[4] A probléma ugyanaz, mint az aggregált adatoknál ismert dominanciaprobléma: adatvédelmi szempontból aggályos olyan értékösszegek publikálása, melyeknél az értékösszeget két adatszolgáltató dominálja, és ezért ők az értékösszeg ismeretében többé-kevésbé pontos becslést adhatnak a másik domináns adatszolgáltató értékére.

nem torzítja az átlagot, viszont torzítja a variancia- és kovarianciabecsléseket (*Brand* [2002]). A zaj véletlenszerűsége miatt az anonimizált változó varianciája a zaj varianciájával haladja meg az anonimizálatlan változó varianciáját:

$$Var\left(x^a\right) = Var\left(x\right) + Var\left(\varepsilon\right).$$

Ha a zaj független az adatbázisban szereplő változóktól, akkor a zajosított változó más változókkal vett kovarianciája várhatóan változatlan marad. Ha viszont a zajosítás mindkét változóra kiterjed, a zajosított változók kovarianciája az eredeti kovariancia és a zaj varianciájának az összege – feltéve, hogy a zajok szórása megegyezik.

A zajosítási eljárás dokumentálása és a zaj varianciájának publikálása lehetővé teszi a torzítatlan becsléseket (*Kim* [1990], *Brand* [2002]). A zaj varianciájának ismeretében a felhasználó a

$$\widehat{Var\left(x\right)} = Var\left(x^a\right) - Var\left(\varepsilon\right) \qquad /10/$$

képlettel becsülheti az eredeti változó varianciáját és a

$$\widehat{Cov\left(yx\right)} = Cov\left(y, x^a\right) - Var\left(\varepsilon\right) \qquad /11/$$

képlettel a zajosított változók kovarianciáját (*Kim* [1990]). A /10/–/11/ képletekkel természetesen a korrelációs együttható is becsülhető. Mivel a számítógépekkel szó szoros értelemben vett véletlenszámokat nem lehet létrehozni, egy konkrét mintában a zaj kismértékben korrelálhat a zajosítatlan változóval, így a /10/–/11/ egyenleten alapuló becslések torzíthatnak. E technikai tökéletlenségből fakadó esetleges torzítások minimalizálhatók a szisztematikus zajosítással (*Evans–Zayatz–Slata* [1996]).[5] Az információveszteség azzal is csökkenthető, ha eljárást csak a felfedhető egyének részmintáján használják (*Fagan–Greenberg* [1988]).

A zajosítás az utólagos randomizáláshoz hasonlóan tehát rendelkezik azzal a kedvező tulajdonsággal, hogy az eljárás paramétereinek – konkrétan a zaj varianciájának – ismeretében a felhasználó torzítatlan becslést tegyen, még akkor is, ha az eredeti helyett csak a zajosított adatbázist használhatja.

[5] Az eljárás lényege, hogy az adatbázist először a zajosításra váró változó szerint sorba rendezzük. Ezután felváltva adunk pozitív és negatív zajt a megfigyelésekhez. A pozitív és negatív értékeket két külön eloszlásból vesszük, melyek várható értékei szimmetrikusak, szórásaik pedig azonosak. Például: a pozitív értékeket generáló eloszlás átlaga 1, szórása 0,2; a negatív értékeket generáló eloszlás átlaga –1, szórása szintén 0,2. Ha a szórás az átlag abszolút értékéhez képest kicsi, akkor a normális eloszlás tulajdonságai miatt csak nagyon ritkán fordulhat elő, hogy a pozitív (illetve negatív) eloszlás a szándékoktól eltérően negatív (illetve pozitív) zajt generál.

1.6. Kerekítés

Az eljárás során az adott változót előre meghatározott szabályok szerint kerekítik, hogy a pontos értékek visszatartásával az alanyok azonosítása nehezebbé, illetve lehetetlenné válik (*Fischetti* [1998]). A kerekítés során nem feltétlenül egész számra, de százasokra, ezresekre vagy akár tízezresekre történő kerekítés is előfordulhat, amennyiben az adatok érzékenysége azt kívánja, illetve a kerekítő algoritmus meghatározott eloszlásnak megfelelően véletlen számokkal is dolgozhat (*Shlomo* [2005]).

A kerekítés felfogható a zajosítás inverzének: a *védelemre szoruló* változó olyan, mintha az *anonimizált* változóhoz hozzáadnánk egy véletlen számot, a zajt. Persze ez a „zaj" nem normális, hanem egyenletes eloszlást követ a $[-h,+h]$ intervallumon, ahol h az *anonimizált* változó nagyságrendjének a fele. Az ezresekre kerekítés például annak az eljárásnak az inverze, hogy az ezresekre kerekített számokhoz a $[-500,500]$ intervallumból véletlenszerűen kiválasztott számot adunk. Ha ez az analógia helyes, a zajosításból fakadó torzításokat definiáló képletek a kerekítésből adódó torzításokat is leírják – feltéve, hogy az anonimizált és az anonimizálatlan változókat felcseréljük az egyes képletekben. Az anonimizált változó és a zaj varianciáinak ismeretében az anonimizálatlan változó varianciájának becslőfüggvénye

$$\widehat{Var(x)} = Var\left(x^a\right) + \frac{h^2}{3},$$

a korrelációs együttható becslőfüggvénye pedig

$$\widehat{\rho^2} = \frac{Var\left(x^a\right)}{Var\left(x^a\right) + h^2/3} r^2,$$

ahol r a korrelációs együttható az anomizált adatbázisban. A képletekben szereplő $h^2/3$ hányados a $[-h,+h]$ intervallumon értelmezett egyenletes eloszlású változó varianciája.

1.7. Újra-mintavételezés

Az újra-mintavételezés (resampling) során a módosítandó változó eredeti értékei szerint sorba rendezzük az adatbázist, majd a változóból almintákat hozunk létre a *bootstrap* vagy a *jackknife* eljárással. Az almintákat szintén sorba rendezzük, majd hozzáfűzzük az eredeti adatbázishoz. Az anonimizált és nyilvánosságra hozható változó az almintákból számolt átlag lesz.

A *bootstrap* eljárás során *n* elemű mintából újabb *n* elemű, előre meghatározott (a mai számítógépes kapacitásokhoz mérten általában magas, minimum 10 000) számú almintát generálunk visszatevéses, véletlen mintavétel segítségével. Az anonimizált változó értéke az *i*edik megfigyelésnél

$$x_i^a = \frac{\sum_{s=1}^{S} x_{is}}{S}$$

lesz, ahol *S* az alminták száma, *s* az alminta sorszáma, x_{is} pedig az *i*-edik megfigyelés az *s* almintában. Mindegyik almintában igaz az, hogy *x* értékei olyan sorrendben követik egymást, mint az eredeti adatbázisban. (Ha tehát *x* eredeti értékeit növekvő sorrendbe állítjuk, akkor ugyanezt kell tenni mindegyik almintában is.)

A bootstrap eljárás nem zárja ki annak lehetőségét, hogy egy adott almintába ugyanaz az érték többször is bekerül, míg más értékek egyáltalán nem kerülnek be. Sőt, elvileg az is előfordulhat – igaz, elenyészően kis valószínűséggel[6] –, hogy egy adott alminta kizárólag egyetlen esetet tartalmazza *n* duplikátumban. E probléma kezelésére alkalmas a *jackknife* eljárás. Ennek során az *n* elemű mintánkból *n* számú, $n-1$ elemszámú almintát generálunk, minden egyes alminta esetében egy tag elhagyásával. Az elhagyott elem lehet minden esetben más és más vagy véletlenszerűen kiválasztott. A hagyományos jacknife az anonimizálás során nem használható további megkötések nélkül, hiszen minden egyes almintában lesz 1 pótlólagos adathiány. Ha a statisztikai szoftver az adathiányt végtelenként értelmezi, akkor az adathiány mindig a növekvő sorrendbe rendezett alminták utolsó megfigyeléséhez tartozik, tehát az utolsó megfigyelésnél adathiányt generálnánk. E probléma elvi megoldása az lehet, ha valamilyen technikával úgy rendezzük növekvő sorrendbe az almintákat, hogy az adathiány egy véletlenszerűen kiválasztott sorba kerüljön.

1.8. Összegzés

Ebben a részben áttekintjük, milyen mértékben torzítják az anonimizálási eljárások a – súlyozott, illetve súlyozatlan – átlag-, szórás- és kovarianciabecsléseket. Az átlagbecsléseket az eljárások döntő többsége torzítatlanul hagyja. A varianciabecslések torzítatlanságát már csak az adatcsere és az esetszámra azonos utólagos randomizálás garantálja. A mikroaggregálás okozta torzítás elvileg kismértékű, a zajosítással és az esetszámra nem azonos utólagos randomizálás védett állományokból pedig torzítatlanul becsülhető a variancia, ha az adatgazda publikálja az anonimizálási eljárás releváns paramétereit. A kovarianciabecsléseket szinte mind-

[6] A szóban forgó valószínűség n^{-n}.

egyik módszer torzítja, ám itt is érvényes az, hogy az adatvédelem során használt releváns paraméter (vagy paraméterek) publikálása lehetővé teszi a felhasználók számára a becslések korrigálását.

Az áttekintett anonimizálási módszerek közül kiemelt szerepet játszik az adatcsere. Egyrészt az adatcsere során használt paraméterek publikálása lehetővé teszi a felhasználók számára a torzítatlan becsléseket. Másrészt az adatgazdák tipikus célja a legtöbbször kategorikus kváziazonosítók (például településkódok) anonimizálása, a kategorikus változók kifinomult védelmére pedig csak az adatcsere – valamint annak továbbfejlesztett változata, az utólagos randomizálás – alkalmas. Végül: elvileg semmi akadálya, hogy adatcserével folytonos változókat is anonimizáljunk – míg a folytonos változók védelmére kidolgozott technikák kategorikus változókra történő alkalmazása nem magától értődő.[7] Érdemes ezért az 1.2. alfejezetben bemutatott adatcserét és annak statisztikai következményeit alaposabban szemügyre venni.

2. Az adatcsere statisztikai következményei: további eredmények

Ebben a fejezetben egyrészt a mérési hibák elméletének kontextusába helyezzük az adatcsere statisztikai következményeire vonatkozó eredményeinket. Másrészt azt vizsgáljuk, hogy a szóban forgó eredmények robusztusak maradnak-e, ha az adatcsere nem teljesen véletlenszerű.

2.1. Az adatcsere okozta torzítás mint mérési hiba

Az előző alfejezetben láttuk, hogy az adatcsere torzítja a kovarianciákat: a torzítás mértéke pedig a

$$Q_x(y) = 1 - \frac{p}{Var(x)} \frac{\overline{y}_{10} - \overline{y}_{01}}{\overline{y}_1 - \overline{y}_0}$$

mennyiség függvénye. Mivel az egyváltozós regressziós becslés egy kovariancia és egy variancia hányadosa, az anonimizált állományból számolt egyváltozós becslés az anonimizálatlan állományból számolt becslés és $Q_x(y)$ szorzata. Ez az eredmény nagyon hasonlít arra, amely a mérési hibák regressziós becslésekre gyakorolt hatásá-

[7] Nem világos például, hogy a kedvező tulajdonságokkal rendelkező zajosítást hogyan lehetne kategorikus változókra alkalmazni, hiszen ekkor a zaj normális eloszlására vonatkozó feltevést módosítani kell. A mikroaggregálás és a kerekítés kategorikus analógiája az átkódolás, melynek statisztikai következményeit nehéz elemezni.

ra vonatkozik. Képzeljük el, hogy x^a nem adatcserével anonimizált, hanem u mérési hibával mért változó! A mérési hibák becslésekre gyakorolt hatása ismert (*Fuller* [1987], *Maddala* [2004]): a mérési hiba elkoptatja a regressziós együttható abszolút nagyságát, mivel

$$\hat{\beta} = \frac{Cov\left(y, x^a\right)}{Var\left(x^a\right)} = \frac{Cov\left(x, y\right)}{Var\left(x\right) + Var\left(u\right)} = \beta \frac{Var\left(x\right)}{Var\left(x\right) + Var\left(u\right)} = \beta R_x. \qquad /12/$$

R_x az anonimizált változó megbízhatósági együtthatója.

A /12/ és az /5/ egyenlet hasonlósága alapján a Q mennyiséget érdemes *relatív megbízhatósági együtthatónak* nevezni. A relatív jelző arra utal, hogy Q értéke függ attól a változótól, amivel kovarianciát számolunk. A megbízhatósági együttható terminus alkalmazása indokolt, mert az adatcsere következményeit tekintve mérési hiba: ahhoz hasonlóan koptatja a regressziós becsléseket. Az adatcserével anonimizált változó relatív megbízhatósága annál nagyobb, minél nagyobb az anonimizálásra váró változó szórása és minél kisebb az adatcserével érintett megfigyelések aránya. Sőt, a regresszióelemzés kontextusában az adatcsere olyan eljárásnak tekinthető, mintha a magyarázóváltozót u mérési hibával mérnénk, a fiktív mérési hiba varianciája pedig az $R = Q$ azonosság alapján:

$$Var\left(u\right) = \frac{pVar\left(x\right)}{Var\left(x\right) - p}.$$

Ahhoz, hogy Q is normalizált legyen és a fiktív mérési hiba varianciája ne lehessen negatív, az indikátorváltozó varianciájának kisebbnek kell lennie a p paraméternél, azaz teljesülnie kell a

$$p \le \overline{x} - \left(\overline{x}\right)^2$$

egyenlőtlenségnek. Ha például az indikátorváltozó eloszlása szimmetrikus, azaz az egyesek és nullák száma azonos, akkor az adatcsere elvileg a teljes mintára is kiterjedhet, a megfigyeléspárok relatív gyakorisága tehát 1/2, mégis az egyenlőtlenség ennek felét szabja meg korlátként.

2.2. Adatcsere-technikák és a torzítás várható mértéke

Az adatcsere kovarianciabecsléseket torzító hatása a

$$Q_x\left(y\right) = 1 - \frac{p}{Var\left(x\right)} \frac{\overline{y}_{10} - \overline{y}_{01}}{\overline{y}_1 - \overline{y}_0}$$

mennyiség függvénye. Az egyszerű kifejtés kedvéért mindeddig azt feltételeztük, hogy a cserepartnereket véletlenszerűen választjuk ki és teljesül az

$$\overline{y}_{10} - \overline{y}_{01} = \overline{y}_1 - \overline{y}_0 \qquad /13/$$

egyenlőség.

A gyakorlatban azonban ez a feltevés nem teljesül szükségszerűen. Egyrészt a véletlenszerű kiválasztás nyilvánvalóan nem garantálja, hogy a /13/ egyenlet minden mintában teljesül. A pontosság elérésének egyik eszköze a rétegzés lehet. Ha a rétegképző ismérvek korrelálnak az *y* változóval, és adatcserére az adott rétegeken belül kerül sor, akkor nagy mintákban és tömeges mértékű adatcserénél /13/ egyenletnek teljesülnie kell. A rétegzésnek viszont az a mellékkövetkezménye, hogy egy adott rétegen belül az adatcsere kevés megfigyelésre terjedhet ki, az esetszám csökkenése viszont veszélyezteti a /13/ egyenlőség fennállását. A rétegzés módszerét *Shlomo–Tudor–Groom* [2010] használták, és célzott adatcserének (targeted data-swapping) nevezték.

A „nem véletlenszerűség" szándékosan is előidézhető. Az egyik az irányítottság – abban az értelemben, hogy az $x = 1$ megfigyeléseket az $x = 0$ megfigyelések egyik részhalmazából választjuk ki. Képzeljük el, hogy szükség van a legmagasabb iskolai végzettség anonimizálására, mert egyes diplomások más ismérvekkel együtt beazonosíthatók. Tegyük fel, hogy a diplomásokat a hozzájuk leginkább hasonló érettségizettekkel akarjuk felcserélni. Ebben az esetben az *x* indikátorváltozó a diplomásokat azonosítja, az $x = 0$ feltétel a nem diplomásokat jelöli. Az adatcsere azonban nem terjedhet ki az $x = 0$ halmaz minden elemére – csak azokra, akik érettségizettek. Ha a kutatás során vizsgált *y* változó korrelál az iskolai végzettséggel, akkor *y* átlaga magasabb az érettségizettek körében, mint az összes nem diplomás körében. Emiatt $\overline{y}_{01} > \overline{y}_0$ és ezért $\overline{y}_{10} - \overline{y}_{01} < \overline{y}_1 - \overline{y}_0$, feltéve, hogy $\overline{y}_{10} > \overline{y}_1$. Ebben a példában a diplomásokat a hozzájuk *y* szempontból leginkább hasonló nem diplomásokra cseréltük. A hasonló megfigyeléseket célzó irányított adatcsere növeli a *Q* megbízhatósági együtthatót.

3. Szimulációs vizsgálatok

Az adatcsere kovarianciabecslésekre gyakorolt hatása – a /13/ egyenletben megfogalmazott feltevés mellett – ismert. Nem világos azonban, hogy az eredmények kiterjeszthetők a többváltozós becslések kontextusára. A többváltozós modellek legkisebb négyzeteken alapuló becslése ugyanis a magyarázóváltozók varianciakovarianciamátrixa inverzének és a függő és a magyarázóváltozók kovarianciamátrixának (pontosabban vektorának) szorzata. A mátrixalgebra miatt nehezen látha-

tó át, milyen mértékben torzulnak a becslések, ha például az egyik magyarázóváltozó varianciája az adatvédelem miatt megnő. A probléma hasonló ahhoz, amikor mérési hiba folytán egy adott változó szórása nő, és ezáltal az összes változó együtthatójának regressziós becslése módosul (*Maddala* [2004]). Az analitikus eredmények hiánya vagy értelmezési nehézségei indokolják a szimulációs módszerek használatát. Az anonimizálási eljárások statisztikai következményeinek szimulációs vizsgálata bevett gyakorlat. A regressziós becslésekre gyakorolt hatások vizsgálata azonban eddig főleg a mikroaggregálásra korlátozódott (*Liu–Little* [2003]; *Lenz et al.* [2006]; *Schmid–Schneeweiss* [2005], [2007], [2008]).

Tekintsük az egyszerű

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

többváltozós modellt. Ha az adatbázis nem szorul védelemre, az együtthatókat az alábbi képletekkel számoljuk ki:

$$\beta_1 = \frac{\dfrac{Cov(yx_1)}{Var(x_1)} - \dfrac{Cov(yx_2)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - \dfrac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}},$$

$$\beta_2 = \frac{\dfrac{Cov(yx_2)}{Var(x_2)} - \dfrac{Cov(yx_1)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - \dfrac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}}.$$

Ha viszont az első magyarázóváltozót adatcserével védik, az anonimizált állományban értelmezett

$$y = \beta_0 + \beta_1 x_1^a + \beta_2 x_2 + \varepsilon$$

modell becsléséhez az alábbi becslőfüggvényeket kell használni:

$$\hat{\beta}_1 = \frac{Q_1(y)\dfrac{Cov(yx_2)}{Var(x_2)} - Q_1(x_2)\dfrac{Cov(yx_1)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - Q_1(x_2)\dfrac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}}, \qquad /14/$$

$$\hat{\beta}_2 = \frac{\frac{Cov(yx_2)}{Var(x_2)} - Q_1(y)Q_1(x_2)\frac{Cov(yx_1)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - Q_1(x_2)\frac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}}.$$

Az anonimizálás mindkét együtthatót érinti, tehát az anonimálás által nem érintett változó együtthatója is torzul. A torzítás irányát és nagyságát nehéz előre jelezni: hiába ismerjük az egyes Q értékeket, a torzítás nagysága a többi kovariancia és variancia nagyságától is függ. A nehézség analóg azzal a problémával, hogy a mérési hibákból fakadó torzításokat is nehéz a vonatkozó képletek alapján előre jelezni (*Maddala* [2004]). A tanulmány hátralevő részében ezért szimulációs módszerrel vizsgáljuk az analitikusan nem előre jelezhető torzítások mértékét.[8]

3.1. Adatok és módszerek

A szimulációs vizsgálattal egy olyan kutató helyzetébe képzeljük magunkat, aki arra kíváncsi, milyen mértékben befolyásolja az iskolai végzettség és a településtípus a munkaerő-piaci aktivitást és a kereseteket. A kutató egy olyan nagymintás felméréshez szeretne hozzáférni, mely e változók mellett a munkaerő-piaci siker alternatív okairól – például az életkorról, a háztartásban élő gyermekek számáról és a nemről – is tartalmaz információkat. Az anonimizálatlan adatokhoz való hozzáférés azonban lehetetlen – feladatunk annak vizsgálata, mennyiben torzítja az adatcserével végzett adatvédelem a kutató becslési eredményeit.

A szimulációhoz a KSH munkaerő-felvétel 2011. első negyedéves adatait használjuk. Az adatbázis valójában már anonimizált; jelen kutatás keretében azonban úgy teszünk, mintha az anonimizálatlan adatbázis lenne birtokunkban. (Az anonimizálatlan adatbázishoz nem férhettünk hozzá.) Az adatbázis 47 162 egyénről tartalmaz adatokat; közülük 23 783 dolgozott a megkérdezés idején. Az iskolázottságot három indikátorváltozóval mérjük: szakmunkás végzettség (ISK_2), érettségi (ISK_3) és diploma (ISK_4). A településtípus (*TELTIP*) kategorikus változó, melynek definíciója: 1 = Budapest, 2 = megyei jogú város, 3 = egyéb város és 4 = község. Az „egyéb város” helyett a továbbiakban kisváros elnevezést használjuk. A nem olyan indikátorváltozó, melynek 1-es értékei a férfiakra vonatkoznak. Használni fogjuk a

[8] Az eredményeket az utólagos randomizálásra is érvényesnek tekintjük. Egy szimulációs vizsgálat kontextusában az adatcsere és az utólagos randomizálás ekvivalens. A gyakorlatban e két technika csak abban tér el, hogy az előbbinél tudatosan, az utóbbinál véletlenszerűen dől el, melyik megfigyeléssel cserélünk fel egy adott megfigyelést. Szimulációs vizsgálatokban azonban csak véletlenszerű cserék léteznek, tudatosan kiválasztott cserék nem.

gyermekek jelenléte indikátorváltozót is, melynek értéke akkor 1, ha van a háztartásban 0–6 éves gyermek.

A munkaerő-felvételben nincsenek béradatok. Pótlásként szimulált jövedelemváltozót hozunk létre a rendelkezésre álló változókra támaszkodva. A szimulált logaritmus jövedelem definíciója:

$$\log \text{ kereset} = 9,67 + 0,1ISK_2 + 0,2ISK_3 + 0,6ISK_4 - 0,1(TELTIP - 1) + \\ + 0,5KOR - 0,0002KOR^2 + 0,2NEM + e,$$

ahol e standard normális eloszlást követő véletlen szám. Az együtthatókat *Kertesi–Köllő* [2001] eredményei inspirálták (lásd az idézett tanulmány F2. táblázatát). A reziduum szórása egységnyi, a determinációs együttható (R^2) értéke így durván 25 százalék.

A fiktív kutató célja tehát egyrészt a logaritmus jövedelem, másrészt a munkaerőpiaci aktivitás modellezése, előrejelzése. A kutató azonban csak az anonimizált adatbázishoz férhet hozzá. Szimulációs vizsgálatunk lényege: a szimulált jövedelmet is tartalmazó munkaerő-felvétel adatbázist anonimizálatlannak tekintjük, a képzeletbeli felhasználónak pedig különböző eljárásokkal anonimizált állományokat bocsátunk rendelkezésre. Ezután azt vizsgáljuk, milyen mértékben térnek el az anonimizált állományokból becsült eredmények az anonimizálatlan(nak tartott) állományokban becsült (valós) eredményektől.

Az anonimizálásra 16 eljárást dolgoztunk ki. Mindegyik eljárásban közös az a feltételezés, hogy az anonimizálatlan(nak tekintett) állományban a diplomás falusi megkérdezettek egy része felfedhető – annak ellenére, hogy az adatbázis már nem tartalmaz olyan kváziazonosítókat, mint például a település neve vagy kódja, a foglalkozás neve vagy kódja. Az egyes eljárások három dimenzióban térnek el.

1. A diplomás falusiak védelmét vagy csak az iskolai végzettség, vagy csak a lakóhely, vagy mindkét ismérv együttes, vagy a két ismérv megosztott anonimizálásával oldjuk meg. Az utóbbi azt jelenti, hogy az azonosíthatónak tekintett egyének véletlenszerűen kiválasztott felénél a diplomás végzettséget, a másik felénél a falusi lakóhely változót anonimizáljuk.

2. A donorokat egyszerű véletlen vagy rétegzett kiválasztással választhatjuk ki. A rétegzett kiválasztásnál adatcserére csak a rétegeken belül kerülhet sor. A rétegeket a nem és a korcsoport kombinációi definiálják.[9]

[9] A korcsoportváltozónak 5 kategóriája van, melyek rendre a 16–25, 26–35, 36–45, 46–55 és 56–65 éves egyéneket azonosítják. A rétegek száma tehát 2 × 5 = 10.

3. A donorok – akár egyszerű, akár rétegzett – kiválasztása lehet irányítatlan vagy irányított: irányítatlan kiválasztásnál bárki lehet donor, aki nem diplomás (illetve nem falusi), míg az irányított kiválasztásnál csak a diplomásokra, illetve falusiakra leginkább hasonlító egyének – tehát az érettségizettek, illetve a kisvárosokban lakók – lehetnek donorok.

Mindegyik módszernél azt feltételeztük, hogy a falusi diplomások ($ISK_4 = 1$ és $TELTIP = 4$) p százaléka felfedhető. A szimuláció során p a 10, 25 és 50 értékeket vette fel. A munkaerő-felvétel mintájában az aktív megkérdezettek durván 6 százaléka falusi diplomás. A három értékkel tehát olyan helyeztet modellezünk, amikor egy adatbázisban a megfigyelések rendre 0,6, 1,5 és 3 százaléka fedhető fel. A kísérletet a 16 módszer és a p paraméter mindegyik kombinációjánál ezer alkalommal ismételtük meg.

A 16 módszer és a p paraméter értékei által definiált anonimizált adatbázisokban különböző becsléseket végzünk, és ezeket összehasonlítjuk az anonimizálatlan(nak tekintett) állományban végzett becsléssel. Az összehasonlításokat relatív torzítás formájában prezentáljuk. Egy adott s statisztika relatív torzítását a következőképpen számoljuk ki. Adott módszer és p paraméter mellett az anonimizálatlan (vagy annak tekintett) adatbázist egy adott eljárással R alkalommal anonimizáljuk (vizsgálatainkban $R = 1000$). Az r-edik replikációban a statisztika értéke s_r. Az anonimizálatlan állományban a statisztika értéke S. A relatív torzítás képlete:

$$s \text{ relatív torzítása} = 100\frac{\sum_{r=1}^{R} s_r - RS}{RS}. \qquad /15/$$

Ha a /13/ képletben megfogalmazott feltétel teljesül, a kovarianciabecslések relatív torzítása a /2/ képlet alapján:

$$\frac{\hat{\beta} - \beta}{\beta} = -\frac{p}{Var(x)}. \qquad /16/$$

A relatív torzítás tehát egyenesen arányos az adatcserével védett megfigyelések arányával. A /14/ egyenlet bonyolultsága miatt a többváltozós regressziós becslések relatív torzítását a /16/ egyenlet segítségével sem lehet előre jelezni.

3.2. Eredmények

A kovarianciabecslések torzulásai. A képzeletbeli kutató célja a jövedelmekben és a munkaerő-piaci aktivitásban mérhető egyenlőtlenségek elemzése. Mivel a több-

változós regressziós becslések kovarianciák és varianciák függvényei, érdemes először a kovarianciabecslések relatív torzításait elemezni. Az 1. és 2. táblázat egyrészt a diplomás iskolai végzettség és a szimulált logaritmus jövedelem, másrészt a diplomás végzettség és a falusi lakóhely indikátorváltozók kovarianciáinak relatív torzításait mutatja.

1. táblázat

A diplomás indikátorváltozó és a szimulált logaritmus jövedelem kovarianciájának relatív torzulásai

Adatcsere módszere	$p = 10$ százalék		$p = 25$ százalék		$p = 50$ százalék	
	Átlag	Szórás	Átlag	Szórás	Átlag	Szórás
Iskolai végzettség anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–1,600	0,906	–4,000	1,365	–8,192	1,799
véletlenszerűen kiválasztott közeli donorokkal	–1,924	1,016	–4,764	1,469	–9,622	1,902
rétegzéssel kiválasztott donorokkal	–2,142	0,770	–5,413	1,159	–11,043	1,596
rétegzéssel kiválasztott közeli donorokkal	–2,089	0,889	–4,991	1,341	–10,188	1,703
Településtípus anonimizálása:	0	0	0	0	0	0
Iskolai végzettség és településtípus együttes anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–1,598	0,892	–4,035	1,385	–8,227	1,831
véletlenszerűen kiválasztott közeli donorokkal	–1,916	1,022	–4,748	1,521	–9,476	1,863
rétegzéssel kiválasztott donorokkal	–2,138	0,764	–5,407	1,210	–10,970	1,610
rétegzéssel kiválasztott közeli donorokkal	–2,081	0,871	–5,045	1,299	–10,168	1,580
Iskolai végzettség és településtípus megosztott anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–0,829	0,647	–1,963	1,026	–3,979	1,343
véletlenszerűen kiválasztott közeli donorokkal	–0,941	0,726	–2,445	1,092	–4,775	1,528
rétegzéssel kiválasztott donorokkal	–1,060	0,562	–2,693	0,866	–5,354	1,204
rétegzéssel kiválasztott közeli donorokkal	–1,051	0,617	–2,488	0,942	–5,130	1,273

Megjegyzés. A kovariancia valós nagysága 0,08. A településtípus anonimizálása nem torzítja a vizsgált kovarianciát.

2. táblázat

A diplomás indikátorváltozó és a falusi településtípus indikátorváltozó kovarianciájának relatív torzításai

Adatcsere módszere	$p = 10$ százalék		$p = 25$ százalék		$p = 50$ százalék	
	Átlag	Szórás	Átlag	Szórás	Átlag	Szórás
Iskolai végzettség anonimizálása:						
véletlenszerűen kiválasztott donorokkal	8,224	0,807	20,672	1,256	41,921	1,819
véletlenszerűen kiválasztott közeli donorokkal	11,731	0,770	29,212	1,177	58,391	1,676
rétegzéssel kiválasztott donorokkal	8,110	0,797	20,366	1,271	41,472	1,786
rétegzéssel kiválasztott közeli donorokkal	11,724	0,787	29,222	1,254	58,614	1,724
Településtípus anonimizálása:						
véletlenszerűen kiválasztott donorokkal	8,219	0,810	20,661	1,266	41,915	1,715
véletlenszerűen kiválasztott közeli donorokkal	15,600	0,684	38,818	1,011	77,630	1,408
rétegzéssel kiválasztott donorokkal	8,091	0,795	20,365	1,284	41,368	1,725
rétegzéssel kiválasztott közeli donorokkal	15,236	0,666	37,990	1,056	75,870	1,401
Iskolai végzettség és településtípus együttes anonimizálása:						
véletlenszerűen kiválasztott donorokkal	7,505	0,951	18,475	1,479	36,024	2,148
véletlenszerűen kiválasztott közeli donorokkal	7,438	1,031	18,044	1,569	34,800	2,159
rétegzéssel kiválasztott donorokkal	7,344	0,948	18,161	1,542	35,216	2,097
rétegzéssel kiválasztott közeli donorokkal	7,157	1,044	17,292	1,612	32,834	2,163
Iskolai végzettség és településtípus megosztott anonimizálása:						
véletlenszerűen kiválasztott donorokkal	8,141	0,791	20,371	1,264	40,957	1,777
véletlenszerűen kiválasztott közeli donorokkal	13,591	0,738	33,829	1,119	67,208	1,587
rétegzéssel kiválasztott donorokkal	8,044	0,810	20,206	1,271	40,392	1,775
rétegzéssel kiválasztott közeli donorokkal	13,439	0,755	33,426	1,166	66,286	1,576

Megjegyzés. A kovariancia valós nagysága –0,031.

Az eredmények megfelelnek annak a várakozásnak, miszerint a torzítás mértéke egyenesen arányos az anonimizált megfigyelléspárok (p) arányával (lásd a /16/ egyenletet). Durván két és félszer akkora torzításokat tapasztalunk a $p = 25$ oszlopokban, mint a $p = 10$ oszlopokban, és kétszer akkora torzítást a $p = 50$ oszlopokban, mint a $p = 25$ oszlopokban. A 2. táblázatban lényegesen nagyobb torzításokat tapasztalunk, mint az 1. táblázatban. A szimulált logaritmus jövedelem és a diploma kovarianciájának torzítását nagyon alacsonyan lehet tartani, ha az anonimizált megfigyelléspárok aránya 10 vagy 25 százalék.

Vegyük ezután szemügyre, hogy az adatcsere melyik módszere minimalizálja a kovarianciabecslések torzítását. A szimulált logaritmus jövedelem és a diplomás in-

dikátorváltozó kovarianciájának torzítását a véletlenszerű és irányítatlan kiválasztás minimalizálja a donorok kiválasztásának négy módszere közül. Talán meglepő, de sem a rétegzés, sem az irányítás – azaz a donorok halmazának szűkítése és a közeli donorok kiválasztása – nem javít az eredményeken. A munkaerő-felvételhez hasonló állományokban a donorok kiválasztásakor tehát érdemes a véletlenre hagyatkozni, és nem érdemes sem tudatos szűkítéssel, sem rétegzéssel a donorok kiválasztásába beavatkozni. A torzítások legkisebb mértékben a rétegzett kiválasztásnál szóródnak, de a szórások között jóval kisebb a különbség, mint az átlagos torzítások között.

A falusi településtípus és a diplomás indikátorváltozó kovarianciájának torzítását ezzel szemben a rétegzett kiválasztás minimalizálja. A véletlenszerű és a rétegzett kiválasztás közötti különbség azonban elenyésző mértékű, és jóval kisebb annál, amit a jövedelem és a diplomás végzettség kovarianciájának az elemzésekor találtunk. Tehát továbbra is fenntartható az a következtetés, hogy a donorok kiválasztásakor a rétegzésnek nincs hozzáadott értéke.

Egyváltozós lineáris regressziós becslések torzulásai. Képzeletbeli kutatónk egyik fő célja a (szimulált) jövedelmi különbségek elemzése. Kutatását a diplomások és érettségizettek, illetve kisvárosiak és falusiak közötti jövedelmi különbségek leírásával kezdi. Az egyváltozós regressziós becslésekben bekövetkező relatív torzításokat a 3. és 4. táblázatok mutatják.

A torzítás mértéke ismét egyenesen arányos a p paraméterrel. A legfontosabb eredmény az, hogy az adatvédelembe bevont megfigyelléspárok arányától függetlenül mindig található olyan eljárás, amely a relatív torzítást 10 százalék alatt tartja. Az adatok alapján a torzítás akkor minimális, ha az adatcserét megosztjuk az iskolai végzettség és a településtípus között, a donorokat pedig véletlenszerűen választjuk ki. Tehát sem a rétegzés, sem az irányítás – azaz a donorok halmazának szűkítése és a közeli donorok kiválasztása – nem javít az eredményeken. Egyik eredmény sem meglepő. A megosztott adatcsere sikere annak tulajdonítható, hogy egy adott változónál az adatcserében részt vevő megfigyelések – és ezáltal a p paraméter effektív nagysága – a felére csökken. A véletlenszerű és irányítatlan kiválasztás sikere pedig annak tudható, hogy az egyváltozós regressziós becslések kovarianciák és varianciák hányadosai, az adatcsere pedig csak az előbbit torzítja.

Többváltozós lineáris regressziós becslések torzulásai. Képzeletbeli kutatónk tudja, hogy az egyváltozós regressziós együtthatók torzan mérik az oksági hatásokat, ezért az egyváltozós elemzések után olyan többváltozós lineáris regressziós modellt becsül, melynek magyarázóváltozói: az iskolai végzettséget mérő három indikátorváltozó, a településtípust mérő három indikátorváltozó, a nem, az életkor és annak négyzete. Az iskolázottságnál és a településtípusnál az alapfokú végzettség, illetve Budapest a referenciakategória. A kutató arra kíváncsi, mennyivel haladja meg a diplomások ceteris paribus keresete az érettségizettét, illetve mennyivel haladja meg a „kisvárosiak" (azaz a nem megyei jogú városok) lakóinak *ceteris paribus* keresete a falusiak keresetét. A többváltozós regressziós becslések relatív torzításait az 5. és 6. táblázatok mutatják.

Az egyváltozós becslések elemzésekor azt találtuk, hogy *1*. a torzítást a véletlenszerű adatcsere minimalizálja, *2*. ha lehet, érdemes a falusi diplomások védelmét megosztott adatcserével biztosítani, és *3*. a torzítás mértéke még viszonylag tömeges adatcsere esetén is 10 százalék alatt tartható. A többváltozós együtthatók relatív torzításaira e következtetések közül csak az első illik maradéktalanul: most is a donorok véletlenszerű és irányítatlan kiválasztása a legjobb módszer. A *2*. következtetés most csak a diplomások kereseti előnyére vonatkozó becslésekre igaz – a kisvárosiak kereseti előnyére vonatkozó becslések torzulásait ezzel szemben az együttes anonimizálás minimalizálja. Végül: a többváltozós együtthatók nagyobb mértékben torzulnak, mint az egyváltozós együtthatók.

3. táblázat

Diplomások és érettségizettek közti nyers bérkülönbség relatív torzítása

Adatcsere módszere	$p = 10$ százalék		$p = 25$ százalék		$p = 50$ százalék	
	Átlag	Szórás	Átlag	Szórás	Átlag	Szórás
Iskolai végzettség anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–1,686	0,926	–4,174	1,483	–8,306	1,931
véletlenszerűen kiválasztott közeli donorokkal	–2,590	1,318	–6,378	2,018	–12,720	2,559
rétegzéssel kiválasztott donorokkal	–2,158	0,849	–5,377	1,278	–10,844	1,737
rétegzéssel kiválasztott közeli donorokkal	–2,632	1,146	–6,797	1,748	–13,522	2,099
Településtípus anonimizálása:	0	0	0	0	0	0
Iskolai végzettség és településtípus együttes anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–1,628	0,960	–4,203	1,483	–8,384	1,941
véletlenszerűen kiválasztott közeli donorokkal	–2,558	1,327	–6,315	1,965	–12,601	2,600
rétegzéssel kiválasztott donorokkal	–2,176	0,852	–5,322	1,310	–10,864	1,707
rétegzéssel kiválasztott közeli donorokkal	–2,647	1,174	–6,832	1,786	–13,642	2,217
Iskolai végzettség és településtípus megosztott anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–0,837	0,705	–2,087	1,068	–4,161	1,501
véletlenszerűen kiválasztott közeli donorokkal	–1,322	0,968	–3,139	1,477	–6,252	1,999
rétegzéssel kiválasztott donorokkal	–1,095	0,593	–2,685	0,971	–5,313	1,289
rétegzéssel kiválasztott közeli donorokkal	–1,298	0,827	–3,450	1,279	–6,859	1,761

Megjegyzés. Bérkülönbségen a szimulált logaritmus bérben mutatkozó különbséget értjük. A bérkülönbség valós nagysága 0,464. A településtípus anonimizálása nem torzítja a vizsgált bérkülönbséget.

4. táblázat

Falusiak és kisvárosiak közti nyers bérkülönbség relatív torzítása

Adatcsere módszere	$p = 10$ százalék		$p = 25$ százalék		$p = 50$ százalék	
	Átlag	Szórás	Átlag	Szórás	Átlag	Szórás
Iskolai végzettség anonimizálása:	0	0	0	0	0	0
Településtípus anonimizálása:						
véletlenszerűen kiválasztott donorokkal	0,903	1,667	2,182	2,625	4,295	3,497
véletlenszerűen kiválasztott közeli donorokkal	2,911	2,597	7,172	3,831	14,065	5,015
rétegzéssel kiválasztott donorokkal	1,689	1,450	4,043	2,303	8,120	3,099
rétegzéssel kiválasztott közeli donorokkal	4,576	2,325	11,639	3,593	22,922	4,496
Iskolai végzettség és településtípus együttes anonimizálása:						
véletlenszerűen kiválasztott donorokkal	0,781	1,636	2,224	2,587	4,215	3,416
véletlenszerűen kiválasztott közeli donorokkal	2,894	2,624	7,257	3,902	14,088	5,124
rétegzéssel kiválasztott donorokkal	1,611	1,484	4,046	2,336	8,012	3,112
rétegzéssel kiválasztott közeli donorokkal	4,533	2,380	11,527	3,530	22,851	4,613
Iskolai végzettség és településtípus megosztott anonimizálása:						
véletlenszerűen kiválasztott donorokkal	0,427	1,137	1,083	1,873	2,019	2,568
véletlenszerűen kiválasztott közeli donorokkal	1,412	1,894	3,763	2,946	7,154	3,971
rétegzéssel kiválasztott donorokkal	0,821	1,036	1,997	1,663	4,112	2,326
rétegzéssel kiválasztott közeli donorokkal	2,267	1,671	5,770	2,573	11,460	3,547

Megjegyzés: Bérkülönbségen a szimulált logaritmus bérben mutatkozó különbséget értjük. A bérkülönbség valós nagysága 0,161. Az iskolai végzettség anonimizálása nem torzítja a vizsgált bérkülönbséget.

5. táblázat

Diplomások és érettségizettek közti korrigált bérkülönbség relatív torzítása

Adatcsere módszere	$p = 10$ százalék		$p = 25$ százalék		$p = 50$ százalék	
	Átlag	Szórás	Átlag	Szórás	Átlag	Szórás
Iskolai végzettség anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–3,081	0,993	–7,647	1,592	–15,190	2,075
véletlenszerűen kiválasztott közeli donorokkal	–4,117	1,351	–10,052	2,026	–19,796	2,685
rétegzéssel kiválasztott donorokkal	–3,021	1,011	–7,504	1,524	–15,098	2,100
rétegzéssel kiválasztott közeli donorokkal	–3,922	1,350	–9,969	2,068	–19,642	2,552
Településtípus anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–0,440	0,105	–1,072	0,185	–2,038	0,301
véletlenszerűen kiválasztott közeli donorokkal	–0,587	0,078	–1,350	0,174	–2,308	0,367
rétegzéssel kiválasztott donorokkal	–0,442	0,113	–1,064	0,192	–2,023	0,287
rétegzéssel kiválasztott közeli donorokkal	–0,580	0,080	–1,336	0,181	–2,271	0,369
Iskolai végzettség és településtípus együttes anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–3,049	1,042	–7,740	1,648	–15,489	2,103
véletlenszerűen kiválasztott közeli donorokkal	–3,800	1,379	–9,432	0,200	–18,642	2,664
rétegzéssel kiválasztott donorokkal	–3,052	1,028	–7,506	1,622	–15,257	2,111
rétegzéssel kiválasztott közeli donorokkal	–3,661	1,402	–9,371	2,147	–18,686	2,686
Iskolai végzettség és településtípus megosztott anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–1,752	0,730	–4,350	1,134	–8,576	1,571
véletlenszerűen kiválasztott közeli donorokkal	–2,398	0,978	–5,688	1,546	–10,975	2,068
rétegzéssel kiválasztott donorokkal	–1,747	0,705	–4,270	1,157	–8,391	1,559
rétegzéssel kiválasztott közeli donorokkal	–2,233	0,970	–5,703	1,519	–11,090	2,099

Megjegyzés: Bérkülönbségen a szimulált logaritmus bérben mutatkozó különbséget értjük. A bérkülönbség valós nagysága 0,4.

6. táblázat

Falusiak és kisvárosiak közti korrigált bérkülönbség relatív torzítása

Adatcsere módszere	$p = 10$ százalék		$p = 25$ százalék		$p = 50$ százalék	
	Átlag	Szórás	Átlag	Szórás	Átlag	Szórás
Iskolai végzettség anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–4,611	0,694	–11,002	1,131	–20,423	1,649
véletlenszerűen kiválasztott közeli donorokkal	–5,836	0,575	–13,539	0,984	–23,925	1,574
rétegzéssel kiválasztott donorokkal	–4,506	0,713	–10,828	1,099	–20,194	1,649
rétegzéssel kiválasztott közeli donorokkal	–5,906	0,566	–13,593	1,003	–24,024	1,487
Településtípus anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–2,490	2,665	–6,229	4,145	–12,746	5,701
véletlenszerűen kiválasztott közeli donorokkal	–4,637	4,119	–11,528	6,049	–23,129	8,183
rétegzéssel kiválasztott donorokkal	–2,437	2,554	–6,480	4,205	–13,083	5,489
rétegzéssel kiválasztott közeli donorokkal	–4,916	4,146	–11,635	6,479	–23,443	8,330
Iskolai végzettség és településtípus együttes anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–1,161	2,814	–1,646	4,513	–1,132	6,004
véletlenszerűen kiválasztott közeli donorokkal	3,965	4,359	10,386	6,481	21,544	8,476
rétegzéssel kiválasztott donorokkal	–1,120	2,841	–2,081	4,556	–1,939	5,951
rétegzéssel kiválasztott közeli donorokkal	3,556	4,400	9,884	6,501	21,085	8,516
Iskolai végzettség és településtípus megosztott anonimizálása:						
véletlenszerűen kiválasztott donorokkal	–3,571	1,917	–8,617	3,074	–16,178	4,235
véletlenszerűen kiválasztott közeli donorokkal	–5,249	2,959	–12,026	4,755	–22,414	6,460
rétegzéssel kiválasztott donorokkal	–3,488	1,921	–8,455	3,043	–15,947	4,325
rétegzéssel kiválasztott közeli donorokkal	–5,344	3,068	–12,420	4,770	–22,749	6,456

Megjegyzés. Bérkülönbségen a szimulált logaritmus bérben mutatkozó különbséget értjük. A bérkülönbség valós nagysága 0,089.

Többváltozós probit regressziós becslések torzulásai. Képzeletbeli kutatónk másik kérdése az, mekkora a munkaerő-piaci részvételben mért előnye a diplomásoknak és a kisvárosiaknak az érettségizettekhez, illetve a falusiakhoz képest. Mivel a munkaerő-piaci részvétel diszkrét, az előrejelzés kézenfekvő módszere a többváltozós

probit regresszió. A magyarázóváltozók: az iskolai végzettséget mérő három indikátorváltozó, a településtípust mérő három indikátorváltozó, a nem, az életkor és annak négyzete, valamint a 6 éves vagy annál fiatalabb gyermek jelenléte a háztartásban.

A lineáris regressziós becslésekkel kapcsolatos eredmények szerint a torzítást akkor minimalizálhatjuk, ha a donorokat véletlenszerűen – tehát irányítás, illetve rétegzés nélkül – választjuk ki. Azt is láttuk, hogy a torzítás leginkább az adatcsere megosztásával mérsékelhető. A lineáris regresszióra vonatkozó kvalitatív következtetések érvényesek a probit modelleknél bekövetkező torzításokra is. Ennek oka az, hogy a probit modell valójában egy olyan lineáris regressziós modellel ekvivalens, melynek folytonos függő változója nem megfigyelhető, és a megfigyelt diszkrét kimenet birtokában csak annyit tudunk, hogy a nem megfigyelhető függő változó nagyobb nullánál vagy sem. Ha a látens függő változó teljes mértékben megfigyelhető lenne, a probit becsléseket úgy számolnánk ki, hogy a lineáris regressziós becsléseket elosztjuk a becsült reziduum szórásával. Probit modelleknél a reziduum szórása sem ismert, mégis, a probit együtthatók lineáris regressziós együtthatók és nem megfigyelhető reziduális szórások hányadosai. Emiatt a lineáris regressziós becslések torzulásaira vonatkozó kvalitatív következtetések a probit becslésekre is érvényesek.

A szimulált jövedelmekre vonatkozó lineáris regressziós becslések alapján viszont nem lehet megjósolni a probit becslések torzításainak nagyságát. A 7. táblázat az előbbi többváltozós probit modell becsléseiben bekövetkező relatív torzításokat mutatja. A táblázat lényegesen egyszerűbb az előzőknél. Az előző vizsgálatban használt 16 módszer helyett most csak kettőt használunk: a donorok véletlenszerű és irányítatlan kiválasztásán alapuló együttes, illetve a megosztott adatcserét.

7. táblázat

Probit együtthatók relatív torzítása

Adatcsere módszere	$p = 10$ százalék		$p = 25$ százalék		$p = 50$ százalék	
	Átlag	Szórás	Átlag	Szórás	Átlag	Szórás
Diploma – érettségi különbség együtthatója:						
Iskolai végzettség és településtípus együttes anonimizálása	–5,215	1,124	–12,893	1,674	–25,867	2,267
Iskolai végzettség és településtípus megosztott anonimizálása	–2,673	0,804	–6,645	1,198	–13,335	1,649
Egyéb város – falu különbség együtthatója:						
Iskolai végzettség és településtípus együttes anonimizálása	1,813	2,520	5,325	3,825	13,799	5,364
Iskolai végzettség és településtípus megosztott anonimizálása	–2,824	1,606	–6,667	2,448	–12,007	3,696

Megjegyzés. Az anonimizálatlan adatbázisban a diploma-érettségi különbség probit együtthatója 0,467, az egyéb város-falu különbség együtthatója 0,086.

A diplomások előnyét mérő probit együtthatók torzulásai durván másfélszeresei a diplomások előnyét mérő többváltozós OLS-együtthatók torzulásainál. A probit együtthatók torzulásai most is a megosztott adatcserénél kisebbek. A kisvárosiak előnyét mérő probit együtthatók értelmezése nehezebb, mivel a torzítás együttes adatcserénél pozitív, megosztott adatcserénél negatív.

4. Következtetések

A tanulmányban áttekintettük a felfedés elleni védelem statisztikai következményeit, és szimulációs módszerrel vizsgáltuk az adatcsere kovariancia- és regressziós becslésekre gyakorolt hatását. Amellett érveltünk, hogy az adatcserének kitüntetett szerepe van a felfedés elleni védelemben. Egyrészt az adatvédelmi jogszabályok leginkább a kváziazonosítók módosítására ösztönzik az adatgazdákat, és ezek a változók legtöbbször kategorikusak. A kategorikus változók anonimizálására alkalmas technikák közül pedig csak az adatcsere (illetve az utólagos randomizálás) az, amely a változók átlagait és szórásait nem, a kovarianciabecsléseket pedig ismert mértékben módosítja. Az adatcsere többváltozós regressziós becslésekre gyakorolt hatását viszont nehéz pontosan megjósolni. Ezért az adatcsere statisztikai következményeinek elemzését szimulációs módszerrel folytattuk.

A szimulációhoz a KSH munkaerő-felvételének 2011. első negyedéves adatait használtuk. Azt vizsgáltuk, az adatcsere egyes technikái milyen mértékben torzítanak egy- és többváltozós lineáris regressziós, valamint többváltozós probit becsléseket. A szimuláció során olyan adatvédelmi problémát elemzünk, amikor a szokásos kváziazonosítók – például a településkód – anonimizálása nem nyújt tökéletes védelmet, és az anonimitás garantálásához még olyan változókat is anonimizálni kell, melyeket regressziós elemzésekben magyarázóváltozóként fogunk használni. Vizsgálatunkban az iskolai végzettséget és a településtípust használtuk: azt feltételeztük, hogy a falusi diplomások egy része felfedhető maradt.

Azt találtuk, hogy a becslések akkor torzulnak minimális mértékben, ha *1*. az adatvédelem terhét megosztják a felfedési kockázatot növelő változók – példánkban az iskolai végzettség és a településtípus – között, és *2*. ha a védelemre szoruló megfigyeléseknél a kiválasztott ismérveket teljesen véletlenszerűen – tehát rétegzés és irányítás nélkül – cseréljük ki alacsony felfedési kockázatú egyének ismérveivel. A megosztott adatcsere természetesen nem mérsékli szükségszerűen a felfedési kockázatot, és lehetséges, hogy az adatszolgáltatók anonimitása csak a változók együttes anonimizálásával érhető el. Eredményeink szerint az együttes anonimizálás sem jár számottevő pótlólagos torzító hatással.

A szimulációs vizsgálatokban konkrét számokkal tudtuk kifejezni, milyen mértékben torzulnak a becslések az adatcserével anonimizált állományokban. Ha a falusi diplomások 25 százaléka szorul további védelemre, a regressziós becslések relatív torzítása az esetek döntő többségében 10 százalék alatt marad. Vajon elhanyagolható vagy jelentős a 10 százalékos torzítás? A survey statisztika irodalma különbséget tesz mintavételi és a nemmintavételi hibák között (*Sarndal–Swensson–Wretman* [1992], *Biemer–Lyberg* [2003]). Nyilvánvaló, hogy az anonimizálásból fakadó torzítás értelmezhető a mérési hibák elméletén belül.[10] Az anonimizáláshoz adatcserét használtunk, amely alulbecsülheti a kovarianciákat – feltéve, hogy a változók között pozitív az összefüggés. A regresszióelemzés kontextusában ugyanezzel a következménnyel – tehát a (pozitív) együtthatók alulbecslésével – járna, ha a változókat mondjuk zajosítással védenénk, hiszen a zajosítás növeli a varianciákat és ezáltal csökkenti a parciális korrelációkat. A magyarázóváltozók szórása inflálódásának következményeit hagyományosan a méréselmélet tárgyalja; az elmélet szerint a regresszióelemzés kontextusában a magyarázóváltozók mérési hibája koptatja a regressziós együtthatókat (*Fuller* [1987], *Maddala* [2004]). Az anonimizálásból fakadó torzítás következményeit tekintve mérési hiba. Valószínű, hogy az adatgyűjtés során keletkező mérési hibák jóval nagyobbak az anonimizálásból fakadó, nem számszerűsíthető mérési hibáknál. Ráadásul az anonimizálás releváns paramétereinek publikálása lehetővé teszi ez utóbbi hibaforrás kiküszöbölését (*Gouweleeuw et al.* [1998], *Shlomo* [2010]). Az anonimizálásból fakadó 10 százalékos relatív torzítás vélhetően sokkal kisebb az adatgyűjtés során keletkező hibáknál. Ám az anonimizálásból fakadó mérési hiba hozzáadódik ezekhez a hibákhoz. Az anonimizálásból fakadó becslési hibát tehát akkor tolerálhatják a felhasználók, ha az anonimizálatlan adatbázisban már eleve kismértékű a mérési hibák szórása.

Következtetéseink természetesen nem tekinthetők véglegesnek, hiszen azok csak egyetlenegy adatbázis szimulációs elemzésén alapulnak. Lehetséges, hogy a munkaerő-felvételtől eltérő adatokon másfajta eredményeket kaptunk volna. Például érdemes lenne más adatokon ellenőrizni annak a következtetésnek az érvényességét, miszerint a rétegzés nem javít a becslési eredmények megbízhatóságán. További kutatásoknak kell tisztázniuk, hogy másfajta adatbázisokon milyen mértékű (lehet) az adatcseréből fakadó torzítás.

Az elméleti következtetések ideiglenes jellege mellett a tanulmánynak van egy gyakorlati szempontból fontos tanulsága: a felhasználók érdekei akkor sérülnek a

[10] A mintavételi hibák elméletének használata nehezebb. Abból indulunk ki, hogy egy adott anonimizálási eljárás alulbecsli a sokasági paramétert. Ezzel párhuzamosan a konfidenciaintervallumokat definiáló alsó és felső értékeket is alulbecsüljük. A mintavételi hibák elmélete azt sugallja, hogy az anonimizálás ugyanolyan következményekkel jár, mint a mintanagyság csökkenése. A 10 százalékos relatív torzítás értelmezéséhez tehát azt a relatív mintanagyság-csökkenést kell kiszámolni, aminek hatására a konfidenciaintervallum alsó határa ugyanolyan mértékben csökken, mint a torzítás hatására. A nehézség abból fakad, hogy a kérdésre adott válasz nem független a sokasági paraméter értékétől.

legkevésbé, ha a felfedés elleni védelmet adatcserével vagy utólagos randomizálással biztosítják az adatgazdák. Ezen technikák egyrészt számos kedvező statisztikai tulajdonsággal rendelkeznek, másrészt eleve a kategorikus kváziazonosítók módosítására szolgálnak – ezek azok a változók, melyek tipikusan lehetővé teszik a felfedést. Az adatcsere és az utólagos randomizálás elterjedését azonban más érdekek és megfontolások korlátozhatják. Egyrészt ezek nem védik mindig tökéletesen a személyes adatokat: az eljárások a valóságban nem létező attribútumkombinációkat hozhatnak létre, melyek ismeretében a rosszindulatú felhasználó rájöhet arra, mely esetek védelme miatt használták a módszert (*Gouweleeuw et al.* [1998], *Boudreau* [2005] 9. old.). Másrészt ezek a technikák kifinomultan ugyan, de „manipulálják" az adatokat, ami megrendítheti az adatokat kezelő intézményekkel szembeni bizalmat (*Evans–Zayatz–Slata* [1996]). Ezzel szemben az olyan egyszerű technikák, mint egyes kváziazonosítók visszatartatása vagy néhány egyszerű eset törlése „manipulációktól" mentes anonimizált állományokat hoznak létre. Az egyszerű törlésnél vagy adatvisszatartásnál kifinomultabb technikák elterjedésére tehát akkor lehet számítani, ha már eleve bíznak a statisztikával foglalkozó intézményekben, vagy ha a felhasználók megértik, és nem hamisításként értékelik az anonimizálási technikákat.

Irodalom

BÁNSZEGI K. [1997]: Felfedést akadályozó módszerek a statisztikai tájékoztatásban. *Statisztikai Szemle*. 75. évf. 12. sz. 1039–1046. old.

BIEMER, P. P. – LYBERG, L. E. [2003]: *Introduction to Survey Quality*. John Wiley & Sons. Hoboken.

BOUDREAU, J.-R. [2005]: *Data Swapping is Not the Panacea.* Proceedings of Statistics Canada's Symposium 2005. "Methodological Challenges for Future Information Needs." Statistics Canada.

BRAND, R. [2002]: Microdata Protection through Noise Addition. In: *Domingo-Ferrer, J.* (ed): *Inference Control in Statistical Databases. From theory to practice*. Springer. Berlin. pp. 97–116.

BYCROFT, C. – MERRETT, K. [2005]: *Experience of Using a Post Randomisation Method at the Office for National Statistics*. Monographs of Official Statistics. UNECE and Eurostat Work Session on Statistical Data Confidentiality. Geneva.

COMMISSION OF EUROPEAN COMMUNITIES [2002]: Comission Regulation (EC) No 831/2002. http://eur-lex.europa.eu/pri/en/oj/dat/2002/l_133/l_13320020518en00070009.pdf

DALENIUS, T. – REISS, S. P. [1982]: Data-swapping. A Technique for Disclosure Control. *Journal of Statistical Planning and Inference.* Vol. 6. No. 1. pp. 73–85.

DOMINGO-FERRER, J. – TORRA, V. [2001a]: Disclosure Control Methods and Information Loss for Microdata. In: *Doyle, P. – Lane, J. – Theeuwes, J. – Zayatz, L.* (eds.): *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. North-Holland, Amsterdam. pp. 93–112.

DOMINGO-FERRER, J. – TORRA, V. (2001b): A Quantitative Comparison of Disclosure Control Methods for Microdata. In: *Doyle, P. – Lane, J. – Theeuwes, J. – Zayatz, L.* (eds.): *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. North-Holland, Amsterdam. pp. 113–134.

ERDEI V. – HORVÁTH R. [2004]: Az adatfelfedés elleni védelem statisztikai eszközei. *Statisztikai Szemle*. 82. évf. 8. sz. 705–727. old.

EVANS, T. – ZAYATZ, L. – SLANTA, J. [1996]: *Using Noise for Disclosure Limitation of Establishment Tabular Data.* US Bureau of the Census. www.census.gov/prod/2/gen/96arc/iiaevans.pdf

FAGAN, W. T. – GREENBERG, B. [1985]: *Algorithms for Making Tables Additive: Raking, Maximum Likelihood, and Minimum Chi-Square*. US Bureau of the Census. www.amstat.org/sections/srms/proceedings/papers/1988_086.pdf

FISCHETTI, M. – SALAZAR, J. J. [1998]: Experiments with Controlled Rounding for Statistical Disclosure Control in Tabular Data with Linear Constraints*. Journal of Official Statistics*. Vol. 14. No. 4. pp. 553–565.

FULLER, W. A. [1987]: *Measurement Error Models*. John Wiley & Sons. New York.

GOUWELEEUW, J. – KOOIMAN, P. – WILLENBORG, L. – DE WOLF, P-P. [1998]: The Post Randomization Method for Protecting Microdata. *Qüestiió.* Vol. 22. No. 1. pp. 145–156.

HUNDEPOOL, A. – DOMINGO-FERRER, J. – FRANCONI, J. – GIESSING, S. – LENZ, R. – NAYLOR, J. – SCHULTE NORDHOLT, E. – SERI, G. – DE WOLF, P.-P. [2010]: *Handbook on Statistical Disclosure Control*. ESSNet SDC. On-line: http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf

KERTESI G. – KÖLLŐ J. [2001]: A gazdasági átalakulás két szakasza és az emberi tőke átértékelődése. *Közgazdasági Szemle*. 48. évf. 11. sz. 897–919. old.

KIM, J. J. [1990]: Subpopulation Estimation for the Masked Data. In: *Proceedings of the ASA Section on Survey Research Methods.* Alexandria. pp. 303–308.

KOOIMAN, P. – WILLENBORG, L. C. R. J. – GOUWELEEUW, J. M. [1997]: *PRAM: A Method for Disclosure Limitation of Microdata*. Research paper 9705. Statistics Netherlands. Voorburg, Heerlen.

LENZ, R. – ROSEMANN, M. – VORGRIMLER, D. – STURM, R. [2006]: *Anonymising Business Micro Data – Results of a German Project.* Statistisches Bundesamt. Berlin.

LIU, F. – LITTLE, R. J. A. [2003]: SMIKe vs. Data Swapping and PRAM for Statistical Disclosure Control in Microdata: A Simulated Study. In: *Proceedings of the ASA Section on Survey Research Methods.* Alexandria. pp- 2497–2502.

MADDALA, G. S. [2004]: *Bevezetés az ökonometriába*. Nemzeti Tankönyvkiadó. Budapest.

MATEO-SANZ, J. M. – DOMINGO-FERRER, J. [1998]. A Comparative Study of Microaggregation Methods. *Qüestiió*. Vol. 22. No. 3. pp. 511–526.

REISS, S. P. [1984]: Practical Data-swapping: The First Steps. *ACM Transactions on Database Systems*. Vol. 9. No. 1. pp. 20–37.

SARNDAL, C. E. – SWENSSON, B. – WRETMAN, J. [1992]: *Model Assisted Survey Sampling*. Springer. New York.

SCHMID, M. [2006]: Estimation of a Linear Model under Microaggregation by Individual Ranking. *Allgemeines Statistisches Archiv*. Vol. 90. No. 3. pp. 419–438.

SCHMID, M. – SCHNEEWEISS, H. [2005]: The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study. In: *Pohlmeier, W. – Ronning, G. – Wagner,*

J. (eds): *E*conometrics of Anonymized Micro Data. *Jahrbücher für Nationalökonomie und Statistik*. Vol. 225. No. 5. pp. 529–543.

SCHMID, M. – SCHNEEWEISS, H. – KUCHENHOFF, H. [2007]: Estimation of a Linear Regression under Microaggregation with the Response Variable as a Sorting Variable. *Statistica Neerlandica*. Vol. 61. No. 4. pp. 407–431.

SCHMID, M. – SCHNEEWEISS, H. [2008]: *Estimation of a Linear Model in Transformed Variables under Microaggregation by Individual Ranking*. Manuscript. University of Munich. Munich.

SCHMID, M. – SCHNEEWEISS, H. [2009]: The Effect of Microaggregation by Individual Ranking on the Estimation of Moments. *Journal of Econometrics*. Vol. 153. No. 2. pp. 174–182.

SHLOMO, N. [2010]: *Measurement Error and Statistical Disclosure Control*. S3RI Methodology Working Papers, M10/05. Southampton Statistical Sciences Research Institute, University of Southampton. Southampton.

SHLOMO, N. – TUDOR, C. – GROOM, P. [2010]: *Data Swapping for Protecting Census Tables*. S3RI Methodology Working Papers, M10/06. Southampton Statistical Sciences Research Institute, University of Southampton. Southampton.

SULLIVAN, C. M. [1992]: *An Overview of Disclosure Principles.* U.S. Bureau of the Census. www.census.gov/srd/papers/pdf/rr92-09.pdf

SZÉP K. – GADÁCSI K. [2010]: *Adatok anonimizálása, hozzáférés a mikroadatokhoz, archiválás*. Előadás a Fényes Elek Műhelyben. Május 26. http://fenyeselekegyesulet.wordpress.com/eloadasok/2010-2/adatok-anonimizalasa-hozzaferes-a-mikroadatokhoz-archivalasi-tevekenysegek/

Summary

The paper reviews the statistical implications of several methods of disclosure control. The effects of data swapping on covariance and linear regression estimates are examined in details. It is argued that data swapping has several advantages over alternative methods. The effect of various data swapping techniques on covariance and linear regression estimates is examined using simulations. The simulation study uses anonymized data from the HCSO Labor Force Survey. The author finds that the relative bias associated with data swapping might be held under 10 percent, and the bias can be minimized by selecting the pairs to be swapped randomly and by manipulating several explanatory variables simultaneously. The results are interpreted within the framework of measurement errors.