AZ ADATHELYETTESÍTÉS MODERN TECHNIKÁJA – „MULTIPLE IMPUTATION (MI)”

Danis Ildikó
Bright Future Humán Kutató és Tanácsadó Kft.

Összefoglaló

A nemzetközi szakirodalomban ma már elvárt metódus, hogy amennyiben mód van a hiányzó adatok kiegészítésére, inkább egy kiegészített, nagyobb elemszámú mintán teszteljük hipotéziseinket, mintsem lecsökkentsük az elemszámot a kérdéses esetek törlésével. Az adathelyettesítés mai legmodernebb módszere a „Multiple Imputation (MI)”, vagyis a Többszörös Helyettesítés. Cikkünkben röviden ismertetjük az eljárás elméleti és módszertani hátterét, és az SPSS programcsomag felhasználó barát moduljának felépítését. Illusztrációként egy kis- és egy nagymintás kutatás tapasztalait is megosztjuk az olvasóval.

Kulcsszavak: adathelyettesítés, Multiple Imputation (MI), SPSS, kutatási tapasztalatok

Az adathelyettesítés klasszikus és modern módszerei

Már az 1970-es évek elejétől alkalmaztak különböző ad hoc eljárásokat a hiányzó adatok kezelésére (ld. hiányzó adattal rendelkező esetek kihagyása az elemzésből, egyszerű helyettesítés csoport átlaggal vagy predikció lineáris regresszió által). A mai napig ezek elterjedt megoldások (még a többváltozós statisztikai programok is ezeket adják meg alapértelmezett metódusként), azonban a modern szimulációs elemzések szerint ezek legtöbbször nem helyénvaló eljárások.

A hiányzó adatokat tartalmazó esetek *egyszerű kihagyása* rendkívül nagy adatvesztési aránnyal járhat. Egy nagyobb, több változót vizsgáló kutatásban számos esetet, és egyben említésre méltó információ mennyiséget lehet veszíteni mindössze soronként egy-két hiányzó adat miatt. Ezzel magyarázó, illetve predikciós modelljeink sérülhetnek. A *csoport átlaggal történő helyettesítés* rombolja a változók eloszlásfüggvényét, konfidencia-intervallumát: megnöveli az eloszlások csúcsosságát, vagyis az átlag értékét több esetben regisztrálhatjuk. Emellett a változók közötti lineáris kapcsolatokat is megváltoztatja, méghozzá a korrelációs együttható közelebb kerül a 0-hoz. Az *egyszerű regressziós eljárásban* két vagy több változó közötti predikciós modell alapján egészítünk ki egy hiányzó adatot. Ez az eljárás az ellenkező irányba torzítja a változók közötti korrelációt: növeli annak értékét. (Ennél már jobb megoldás, ha a prediktált változó mellé egy random reziduálist is kalkulálunk.)

Az egyszerű helyettesítésekkel mindenképpen az a probléma, hogy ha nincs a helyettesített adat mellé hibaérték kalkulálva, a későbbi modellünk romlik, mivel az nem tükrözi a hiányzó adatok bizonytalanságát (ld. túl szűk konfidencia intervallumok, I. típusú hiba aránya nő). A probléma egyre fokozódik, ahogy a változók, illetve a hiányzó információk aránya nő.

Az 1980-as évektől kezdődően (Dempster et al., 1977) elterjedtek a *maximum likelihood alapú EM (Expectation-Maximization)* algoritmuson alapuló helyettesítési technikák, majd az 1990-es évektől (Rubin, 1987; Schafer, 1997) az ún. „*multiple imputation" Bayes-i alapú procedúrák*, amelyeket a mai napig fejlesztenek a különböző statisztikai problémák megoldására. Jelenleg ezt a két eljárástípust ajánlják a hiányzó adatok kezelésére. Ezekben az eljárásokban a hiányzó adatok helyettesítésénél több célt és kritériumot fogalmazhatunk meg. Mivel a kiegészített adatokkal végzett statisztikai analízisek révén megbízható és eredményes következtetéseket kell levonnunk a populációra, illetve az adott mintára nézve, meg kell őriznünk a megfigyelt változók eloszlását és asszociációit. A hiányzó adataink becslésénél kismértékű hibára számítunk, miközben kezelni kívánjuk az adatok bizonytalanságát. A hiányzó adatokra vonatkozó becslésekkel kiegészített változók konfidencia intervalluma 95%-ban kell, hogy fedje a „valós" értékeket. Ha a lefedettség pontos, akkor az I. fajú hiba előfordulási valószínűsége is helyes: 5%. Emellett a konfidencia intervallumokat kellően szűknek várjuk, mert ezzel a II. fajú hibák lehetőségei csökkennek.

A HIÁNYZÓ ADATOK FAJTÁI

A nem-válaszolásnak két fajtája van: az *esetre* és az *itemre* vonatkozó nem-válaszolás. Az elsőnél egy adott személy különböző okok miatt (pl. nem lehetett elérni, megtagadta a részvételt) nem ad válaszokat a teljes változólistára, a másiknál viszont csak egy-két változóra nem érkezik válasz. Kedvező feltételezés, hogy a longitudinális vizsgálatok különböző hullámainál megfigyelhető adathiányok jól prediktálhatók a többi hullámból származó adatokkal.

A hiányzó adatok mechanizmusaként Rubin leírása óta (1987) elkülönítenek három típust: MCAR, MAR és MNAR feltételezéseket. Az *MCAR (missing completely at random)* esetében a hiányok valószínűsége egyáltalán nem függ össze az adatainkkal, ilyenkor a nem-válaszolók olyanok, mint egy random alcsoport. Ez nagyon kevés esetben fordul elő. A *MAR (missing at random)* modelleknél a hiányok valószínűsége csak a megfigyelt egyéb adatoktól függ, de nem a helyettesítendő hiányzótól. Ez a standard feltételezés a legtöbb modern hiányzó adatokat kezelő eljárásnál. Egy sokkal kevésbé megoldható probléma az *MNAR (missing not at random)* helyzet, amikor a hiányzó adat előfordulása pont a hiányzó adat minőségével vagy jelentésével függ össze (bővebben ld. az idézett irodalmakat). A maximum-likelihood módszerek elvárása a MAR helyzet, míg a multiple imputation technikák többnyire már az MNAR problémákat is jól kezelik. (A fenti összefoglalást ld. többek között Schafer és Olsen, 1998, Schafer, 1999; Schafer és Graham, 2002 alapján).

A többszörös helyettesítés módszertana

Az MI (multiple imputation, Rubin, 1987; Schafer, 1997, 1999; Schafer és Olsen, 1998; Schafer és Graham, 2002) szimuláción és legtöbbször Bayes-i alapokon álló technika, ahol a megfigyelt adatokból *m> 1* verzióban modelleznek lehetséges adatokat a hiányzók helyére, majd a végén a Rubin (1987) által ismertetett algoritmus szerint kombinálják az eredményeket (a becsléseket és a szórásokat). A módszer már kisszámú *m* esetén is hatékony: annak függvényében, hogy az adatok hány százaléka hiányzik: 3-tól 20-ig terjedő *m* elégséges egy eredményes modellhez. Érdekes módon nagyarányú hiányzó adatokat is eredményesen kezel a módszer. Általános szabályként olyan változók esetében használhatjuk az imputálást, ahol változónként maximum az adatok 30–40%-a hiányzik, de a teljes adatbázisban nincs több hiányzó, mint a teljes mátrix 10–15%-a. Ezek az arányok a szakirodalom szerint egyáltalán nem adnak okot aggodalomra a helyettesítés metódusát illetően. Az MI célja, hogy a helyettesítésekkel együtt megtartsuk a változók eloszlását és a változók közötti asszociációkat. Az MI elvégzésére több szoftver[1] áll rendelkezésre, de a leginkább felhasználóbarát eszköztárat ma már az SPSS utóbbi verziói nyújtják, amelyek a klasszikusabb EM módszer mellett már tartalmazzák az MI opciókat is.

Egy kismintás longitudinális kutatás tapasztalatai

Évekkel ezelőtt (2006–2007) először *NORM eljárással* dolgoztunk a Budapesti Családvizsgálat (Gervai, 2005) longitudinális adataival, ahol a minta elemszám 103 volt. Ekkoriban még nem álltak rendelkezésre az SPSS programcsomag MI moduljai.

Az imputálás utáni változókészlet leíró paramétereit vizsgálva a változók középértékei és szóródás mutatói nagyon hasonlóak voltak a megfigyelt változók paramétereihez (Danis, 2008). Az egész elemzési folyamat során az eredményeket a *„case deletion / eset kihagyás”* módszerével is ellenőriztük, amikor is csak azoknak az eseteknek a bevonásával végeztük el a statisztikai próbákat, akik az adott változók mindegyikére választ adtak. Az adat-imputálással kiegészített adatbázison végzett elemzések eredményei a számítások típusától függően adtak hasonló vagy kevésbé hasonló eredményeket a nem kiegészített adatbázis eredményeihez képest. A *korrelációs elemzések* szinte egyáltalán nem, csupán pár századnyi különbséget mutattak a két adatbázisban. Ez nyilván az imputálás metódusából is fakad, hiszen annak egyik célja, hogy a kapcsolatokat megtartsa. Legtöbbször némileg (elhanyagolható mértékben ugyan, de) szigorúbb, és ezáltal akár megbízhatóbbnak gondolt eredményeket adott a módszer. Néhány *predikciós modell* (lineáris regressziók) esetében bár a tendenciák hasonlóak voltak, a nem imputált adatbázisban több esetben nem érték el a szignifikanciát azok a számítások, amelyek a kiegészített adatbázisban biztos eredményeket nyújtottak. Legnagyobb eltéréssel a *többváltozós elemzések* esetében találkozhattunk, mivel ezekben az esetekben akár 20–45

[1] Pl. szabadon letölthetők az Internetről: AMELIA, WINMICE, NORM (Schafer és Olsen, 1999)

esetszám különbséggel dolgoztunk a random adathiányok miatt (a teljes minta 103 fő volt). Ezekben az esetekben szinte természetesnek vélhetjük a különbségeket.

TÖBBSZÖRÖS IMPUTÁLÁS SPSS PROGRAMON

Az azóta megjelent frissebb SPSS programcsomagok MI modulja (SPSS Missing Values 17.0) nagyon felhasználóbarát, melynek opcióit a következőkben röviden összefoglaljuk.

Az analízisbe kerülő változók adattípusa lehet nominális, ordinális és metrikus skála alapú is. Az adat-imputálás előtti feltáró elemzés elengedhetetlen része, hogy feltérképezzük a hiányzó adataink sajátosságait. A program pontos képet nyújt a hiányzó adatok mintázatáról: azoknak a változóknak, eseteknek és önálló adatértékeknek az előfordulási gyakoriságáról és arányáról, amelyekben egy vagy több hiányzó érték van. Ezek alapján dönthetünk arról, hogy a mátrix összességében alkalmas-e a helyettesítésre, illetve vannak-e olyan változók, amelyeket a túl sok hiányzó adat miatt ki kell hagynunk a modellből. Az elemző ezek után kiválaszthatja a teljes változókészletből azokat a változókat, amelyek alapján az imputálás modelljét fel szeretné állítani. A döntést vagy az elméleti modellünk, vagy pedig a hiányzó adatok sajátosságai alapján kell meghozni.

A bevonás után kijelölhető, hogy mely változók legyenek prediktorok és melyek magyarázott változók: alapértelmezésben minden változó mindkét sajátsággal bír, de ez az elméleti modellünk alapján megintcsak változtatható. Meghatározhatjuk, hogy hány imputálási fordulót szeretnénk végrehajtatni (alapértelmezettként az m=5 imputálási szett van beállítva), minél több a hiányzó adatok aránya, annál több fordulóra lehet szükség.

Az imputálás módszere *(Imputation Method)* beállításnál érdemes az *„Automatic”* opciót használnunk, mivel ekkor a program végigpásztázza a mátrix adatait, és azok mintázata alapján választja ki a megfelelő módszereket. Legtöbb esetben *MCMC (Markov chain Monta Carlo) modellt* fog alkalmazni a program, ahol az egyes változók értékeinél a többi modellváltozó predikcióit fogja felhasználni bizonyos iterációs szám mellett. Az iterációk száma alapesetben 10, de néha szükség lehet a lépések számát emelni, ha a modell nem konvergál.

Megadhatjuk, hogy változónként milyen minimum-maximum értékeket engedünk meg a hiányzó adatok helyettesítésénél, és azokat hány tizedesjegyre kerekítse a program. Így nem kell számítanunk utólagos outlier-problémákra, és nem kell manuálisan a megfelelő formátumra hozni a rengeteg új adatot, mint korábbi programokban. Emellett egy egyszerű utasítással (a pontos % mint határérték megadásával) kihagyhatjuk azokat a változókat, amelyek túl sok hiányzó adatot tartalmaznak.

Az imputálások utáni adatmátrixokat az eredetivel az élen kérésünkre egy fájlba szerkeszti a program, így akár pontosan leellenőrizhetjük, hogy mely hiányzó adatok helyére milyen értékeket helyettesített a program az egyes szettekben. A választott statisztikai elemzésekben egymás alatti táblázatokban kapjuk meg az eredeti adatbázis és az egyes helyettesített szettek szerint számított eredményeket, majd végül egy összesített, *„pooled” számolást,* amely az összes szett információi után kalkulálódik. A felhasználóbarát output fájlokban azonnal képesek vagyunk összehasonlítani az eredeti adatbázisunk alapján számított eredményeket az imputálás segítségével nyert eredményekkel.

Jelenleg nagy adatbázison (a Heim Pál Kórház 1164 fős mintájában; l. Scheuring és mtsai, 2011) használjuk az SPSS programcsomag MI modulját bíztató eredményekkel. A leíró statisztikák ellenőrzésekor minimális eltéréseket kaptunk átlag- és szórásértékekben, viszont kiküszöböltük azt a heterogén adathiányokból adódó következményt, hogy később nagyon lecsökkenjenek majd a többváltozós elemzések mintaelemszámai. Részletes eredményeinkről, amelyekben az egyes statisztikai próbák eredményei közötti különbségeket is tárgyaljuk majd, a közeli jövőben számolunk be.

Összefoglalás

Azokban a kutatásokban, amelyekben korrelációalapú számításokat végeznek a kutatók, biztonsággal alkalmazható az adat-imputálás. Saját longitudinális elemzéseinkben (Danis, 2008) a különböző magyarázó modellek többé-kevésbé megerősítődtek imputálás nélkül is, imputálás után legtöbb esetben az elemszám különbségekből adódhattak eltérő – kiemelten a szignifikancia szintjét érintő – eredmények, ezért a kiegészített adatbázist tekintettük mérvadónak. Azonban a kutatásokban mindenképpen törekedni kell a minél teljesebb adatbázis létrehozására, eredményeinket ekkor fogadhatjuk el minden fajta szkepszis nélkül.

Summary

The modern method of data imputation: „multiple imputation" (MI)

Imputing the missing data of our sample is a desirable method according to the literature. If we have some solutions for imputation of the missing data we should rather use imputed datasets for testing our hypotheses, than decrease the sample size because of deletions of the problematic cases. The most modern method for data imputation nowdays is „Multiple Imputation (MI)". The theoretic and methodological background of the method and the user-friendly module of the SPSS programme package is introduced shortly in our article. As an illustration, a small and a large sample example is shared with the readers, as well.

Keywords: Data imputation, Multiple Imputation (MI), SPSS, research examples

Irodalom

Danis I. (2008): *Szülői és tágabb környezeti tényezők szerepe a szülővé válás folyamatában és a korai anya-gyermek kapcsolat kialakulásában.* Doktori értekezés. Eötvös Loránd Tudományegyetem, Pedagógiai-Pszichológiai Kar, Pszichológiai Doktori Iskola, Kognitív Fejlődés Program.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. (1977): Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39.* 1–38.

GERVAI, J. (2005): A Budapesti Családvizsgálat. *Alkalmazott Pszichológia, 7.* 5–13.

RUBIN, D.B. (1987): *Multiple imputation for nonresponse in surveys.* New York: Wiley.

SCHAFER, J. L. (1997): *Analysis of incomplete multivariate data.* London: Chapman & Hall.

SCHAFER, J. L. (1999): Multiple imputation: a primer. *Statistical Methods in Medical Research, 8.* 3–15.

SCHAFER, J. L. (2003): Multiple imputation in multivariate problems where the imputer's and analyst's models differ. *Statistica Neerlandica, 57.* 19–35.

SCHAFER, J. L., GRAHAM, J. W. (2002): Missing data: our view of the state of the art. *Psychological Methods, 7.* 147–177.

SCHAFER, J. L., OLSEN, M. K. (1998): Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research, 33.* 545–571.

SCHAFER, J. L., OLSEN, M. K. (1999): *NORM Version 2.02 for Windows 95/98/NT.*

SCHEURING, N., PAPP, E., DANIS, I., NÉMETH, T., CZINNER, A. (2011): A csecsemő- és kisgyermekkori regulációs zavarok háttere és diagnosztikai kérdései. *Gyermekorvos Továbbképzés, X (5)*

SPSS Missing Values 17.0. Manual. SPSS Inc.