Adatpótlás, imputáció Bayes-i módszerekkel

Priksz Tamás[1] , Lang Zsolt[1], Rakonczai Pál[1], Bacskai Miklós[1]
[1]Healthware Tanácsadó Kft.

Háttér - a probléma bemutatása

Probléma felvetés

- Hiányzó értékek: Gyakori probléma statisztikai elemzések során, hogy az adatbázisban található mérési értékeket tartalmazó cellák hiányos kitöltöttségűek. Különböző típusú méréseket tartalmazó adatbázisok egyesítésekor, amikor az elérhető adatok együttes felhasználása a cél, ez a probléma fokozottan jelentkezik.
- Imputáció: A hiányosan kitöltött változó értékeit azonban egy vagy több másik (nem hiányos) változót felhasználva, a hiányos és az elérhető változók közötti kapcsolatok ismeretének segítségével pótolni lehet. Ezt a statisztikai eljárást imputációnak nevezzük. [1]
- Előnyök: Az imputáció eredményeképp az adatbázis szélesebb köre válik elemezhetővé, ezáltal csökkenthető a becslések szórása és növelhető a vizsgálatok statisztikai ereje.

Adatok:

A példánk során rheumatoid arthritisben (RA) szenvedő betegek állapotát jellemző mérőszámok (lsd. *1.* táblázat és *1.* ábra) közül az alábbiakat használjuk fel (*1427* mérés):

- THR – Trombociták száma – egy egészséges ember normális tartománya: *150-400*.
- FVS – A fehérvérsejtek száma a szervezetben lévő gyulladást jelzi – egy egészséges ember normális tartománya: *4,8-10,8*.
- DAS28 – A betegség aktivitási index *28* ízület vizsgálatával – tartománya: *0-10*.
- VAS – Vizuális-analóg fájdalom skála – fájdalom intenzitás tartománya: *0-100*.

A felsorolt jellemzők közötti összefüggések nagyságát a *2.* táblázat tartalmazza.

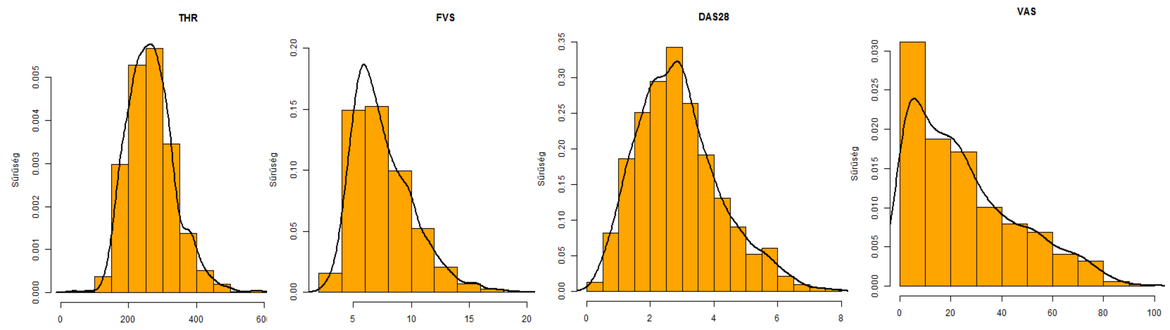
1. táblázat: Leíró statisztika

	Min.	Medián	Átlag	Max.
THR	*28*	*261*	*266,7*	*713*
FVS	*2,4*	*7*	*7,6*	*23*
DAS28	*0,2*	*2,8*	*2,9*	*7,9*
VAS	*0*	*21*	*25,2*	*100*

2. táblázat: Páronkénti lineáris korrelációk

Korrelációk	THR	FVS	DAS28	VAS
THR	*1*	*0,45*	*0,24*	*0,13*
FVS	*0,45*	*1*	*0,19*	*0,17*
DAS28	*0,24*	*0,19*	*1*	*0,73*
VAS	*0,13*	*0,17*	*0,73*	*1*

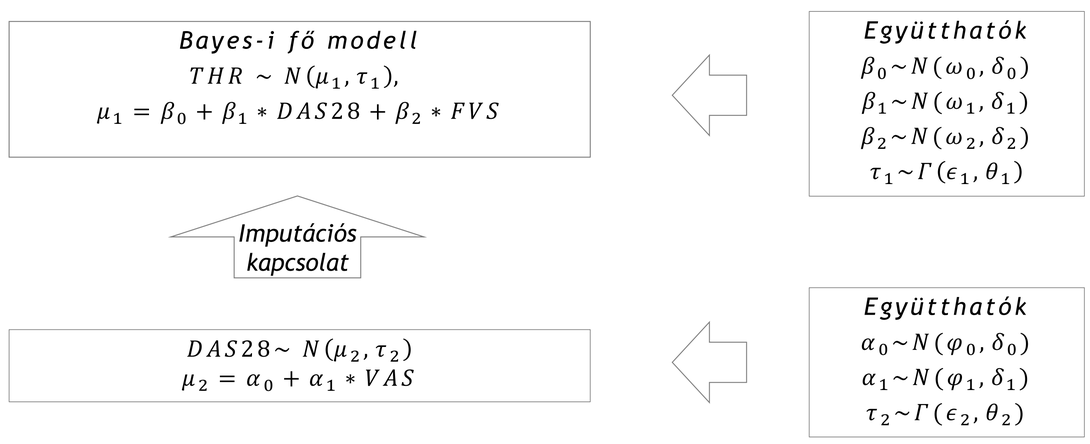
1. ábra: Hisztogramok



Módszertan - a modell felépítése

- **Összefüggés vizsgálat:** A példában a trombociták számának (THR) a fehérvérsejt számtól (FVS) és a DAS28 indextől (DAS28) való függését vizsgáltuk regressziós modellek együtthatóinak becslésének segítségével.
- **Frekventista vs. Bayes-i módszer:** A regressziós együtthatók becslését a frekventista megközelítés mellett (ún. maximum likelihood módszer) Bayes-i megközelítést alkalmazva is elvégeztük. A standard frekventista módszert követve az adatbázis csak azon részére lehet modellt illeszteni, ahol a modell minden változója egyszerre ismert, a Bayes-i módszer esetén van lehetőség imputáció segítségével kiterjeszteni az adatkört. [2] [3]
- **Szimulációs vizsgálat:** Az imputáció hatékonyságának vizsgálatát úgy hajtottuk végre, hogy DAS28 magyarázó változóban fokozatosan növeltük a hiányzó, imputálandó értékek arányát, majd a hiányzó értékeket egy DAS28 változóval jól korreláló (ρ=*0,73*) VAS helyettesítő változó felhasználásával imputáltuk. A modell összetevőit a *2.* ábra illusztrálja. Adott hiányzó érték arányok mellett a csökkent méretű adatbázisokon, illetve azoknak imputált változatán is elvégeztük a együtthatók Bayes-i módszerrel történő becslést. Minden rögzített arány esetén az ismertetett algoritmust az adatok bootstrap szimulációja segítségével *100* alkalommal megismételtünk, majd értékeltük a kapott eredményeket.
- **Algoritmus és szoftverek:** Az összefüggések modellezéséhez az ún. Markov chain Monte Carlo (MCMC) szimulációs algoritmust [4] alkalmazzuk. Az illesztés kivitelezésekor az R statisztikai szoftvercsomagot [5] a JAGS szoftverrel [6] kiegészítve használtuk.

2. ábra: Bayes-i modellek felépítése

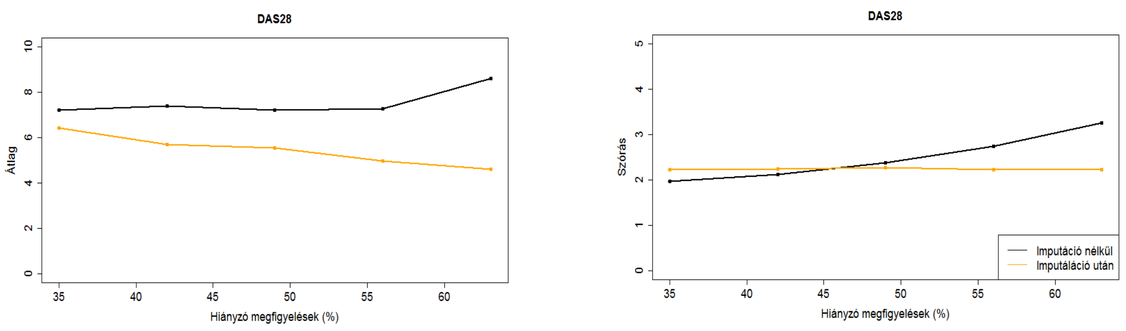


Eredmények - az imputáció hatása

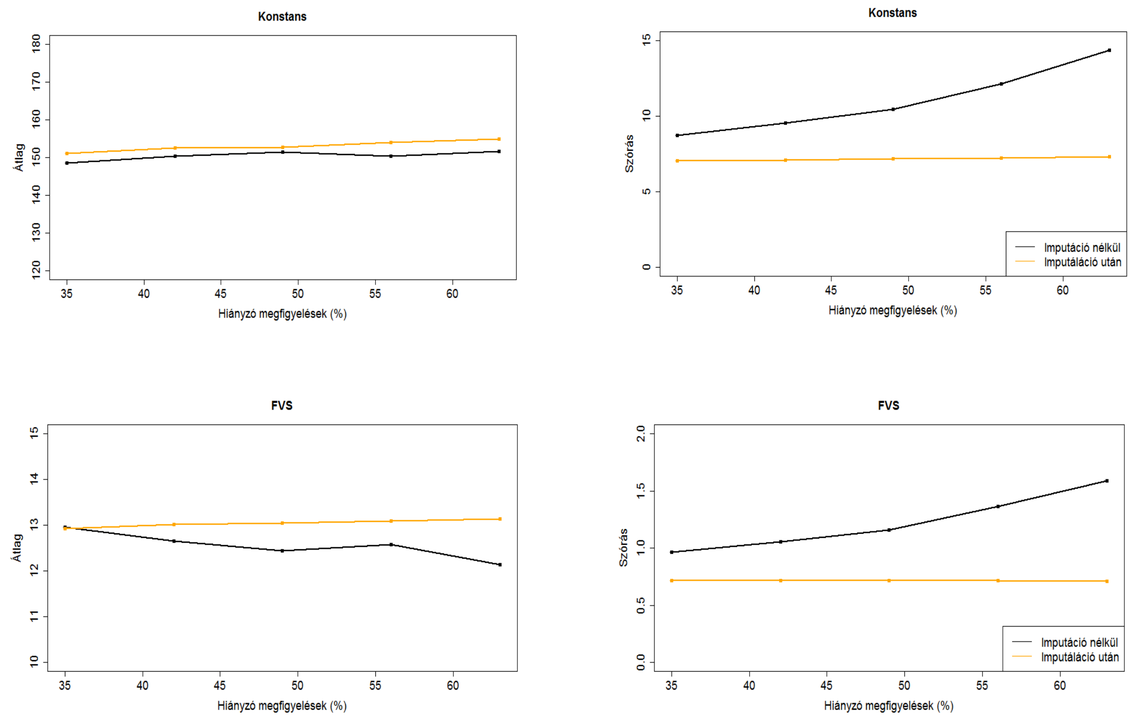
Az *3-4.* ábrán a *2.* ábra modelljének *100* bootstrap mintán számolt Bayes-i becléssel kapott együtthatók becsléseire (β_0, β_1 *és* β_2) vonatkozó statisztikákat (átlag és szórás) mutatjuk be.

- **Elhagyás után végzett becslések:** Az ábrákról látható, hogy a DAS28 mérések elhagyása a DAS28 (β_1) (*3.* ábra, bal oldal, fekete szín) együtthatójának becslését *55%*-os szintig csak kis mértékben befolyásolja, viszont az elhagyással párhuzamosan a becslések szórása folyamatosan nő (*3.* ábra, jobb oldal, fekete szín). Az elhagyás hatása hasonlóan hatott a konstans (β_0) és a fehérvérsejt (β_2) együtthatóinak becsléseire, az átlagok és a szórások tekintetében egyaránt.
- **Imputáció utáni becslések:** Az bootstrap átlagok az imputációt követően (narancs szín) az elhagyás mértékétől függően az elhagyásos esethez hasonlóan változnak. Megállapítható ugyanakkor, hogy a változás mértéke a konstans (β_0) és a fehérvérsejt (β_2) együtthatók becslései esetén kisebb, mint a hiányzó értékeket egyre nagyobb arányban tartalmazó DAS28 (β_1) együtthatójának esetében. Mivel az imputáció következtében az elérhető megfigyelések száma nem csökken, a magasabb elemszámnak köszönhetően az imputált modellek becsléseinek szórása stabilnak mondható. Emellett azt tapasztaljuk, hogy a szórás a vizsgált elhagyási ráták esetén alacsonyabb is.
- **Konklúzió:** A felhasznált adatokon azt tapasztaltuk, hogy a Bayes-i imputáció kedvező hatással van a becsült modell együttható-becsléseinek szórására. Bár ez a hatás kis mértékben jelentkezik magára a hiányzó értékeket tartalmazó változóra nézve (*3.* ábra), az imputáció a modell további változóira - a becslés várható értékét és standard hibáját is figyelembe véve – egyértelmű és erős stabilizáló hatással bír (*4.* ábra).

3. ábra: Bootstrap szimuláció eredményei (β_1 becslés átlaga és szórása) különböző hiányzó érték arányok esetén



4. ábra: Bootstrap szimuláció eredményei (β_0, β_2 becslések átlaga és szórása) különböző hiányzó érték arányok esetén



Hivatkozások

1. A. Gelman, J. Hill: *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2006
2. A. Gelman, J. B., Carlin, H. S., Stern, D. B., Rubin: *Bayesian Data Analysis,* Chapman & Hall/CRC, 2013
3. E.L. Lehmann, G. Casella: *Theory of Point Estimation*, Springer, 2003
4. S., Brooks, A. Gelman, G. L. J. & Xiao-Li Meng (szerk.,): *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, 2011
5. R: https://www.r-project.org/
6. JAGS: http://mcmc-jags.sourceforge.net/

A posztert a Healthware Tanácsadó Kft. készítette.
IME-META X. Egészség-gazdaságtani Továbbképzés és Konferencia, Budapest, 2016. Június 15-16.
Healthware Tanácsadó Kft. / H-1093 Budapest, Közraktár utca 32. River Park Irodák, VII. emelet
Phone.: +36 1 279 1083 / Fax: +36 1 279 1084 / www.healthware.hu

Töltse le poszterünket honlapunkról:

Poszter kód: P24