Bene Márton

ADATELEMZÉS AZ R-BEN

AZ INFORMATIKA ALKALMAZÁSAI

Bene Márton

ADATELEMZÉS AZ R-BEN

Bevezetés a társadalomtudományi adatok elemzésébe az R-program használatával

TYPOTEX

A könyv megjelenését a Magyar Tudományos Akadémia támogatta.

A könyv elkészítését a szerző az Európai Unió támogatásával valósította meg,
az RRF-2.3.1-21-2022-00004 azonosítójú *Mesterséges Intelligencia Nemzeti Laboratórium*
projekt keretében.

Lektorálta Papp Zsófia

ISBN 978 963 493 288 8
ISSN 1787-6044

Kedves Olvasó!
Köszönjük, hogy kínálatunkból választott olvasnivalót!
Újabb kiadványainkról és akcióinkról a *www.typotex.hu*
és a *facebook.com/typotexkiado* oldalakon értesülhet.

Typotex Kiadó
Alapította Votisky Zsuzsa, 1989
A kiadó az 1795-ben alapított Magyar Könyvkiadók
és Könyvterjesztők Egyesülésének tagja.
Felelős kiadó: Németh Kinga
Felelős szerkesztő: Erő Zsuzsanna
Borítóterv: Szalay Éva
Tördelte: Madarász György
Készült a Multiszolg Bt. nyomdájában
Felelős vezető: Kajtor Bálint

TARTALOM

BEVEZETÉS

A tankönyv az R-programban történő társadalomtudományi adatelemzés mikéntjébe vezeti be az olvasót. Az elmúlt években az R vált a társadalomtudományi elemzések elsődleges platformjává, vonzerejét többek között annak is köszönheti, hogy más adatelemző programokkal ellentétben (pl. SPSS, Stata) ingyenesen elérhető. Ráadásul az R nemcsak hagyományos kvantitatív adatelemzésre használható, hanem számos kurrens módszertani eljárás alkalmazására (pl. hálózatelemzés, szövegbányászat, mesterséges intelligenciára építő modellek) is lehetőség nyílik, ha a program alapjait ismerjük. Magyar nyelven azonban átfogó, kifejezetten társadalomtudományi szempontú bevezető könyv ez idáig nem készült a témában, annak ellenére, hogy a specifikusabb felhasználásról, így a szövegbányászati és mesterséges intelligenciára épülő alkalmazásról a közelmúltban jelent meg tankönyv.[1]

A tankönyv részletesen átveszi a társadalomtudományi kvantitatív alapelemzés alapjait, így az adatbázis-kezelés, az egyváltozós, a kétváltozós és a többváltozós elemzések főbb szempontjait. Ezek R-ben való végrehajtását számos példán keresztül mutatja be, részletesen elmagyarázva az egyes parancsok alkalmazását. A parancsok tekintetében a könyv az egyszerűségre törekszik: kerüli az olyan bonyolultabb funkciókat, amelyek bár hosszú távon megkönnyíthetik a parancsok írását, rövid távon nehezítik a megértést és elrettentik az olvasót a program további használatától. Arra törekszik, hogy bemutassa: bár az R-ben kódokból álló parancsokat használunk, néhány egyszerű szabály ismeretében ezek könnyen alkalmazhatóak bármilyen problémára, nem szükséges a társadalomtudományi hátterű érdeklődőnek programozóvá válnia ahhoz, hogy az R-t hatékonyan tudja adatelemzésre használni. A tankönyv olyanoknak is szól, akiknek nincsenek előzetes adatelemzési ismereteik, ezért az elemzések R-ben való elvégzése mellett az egyes elemzési eljárásokról gyakorlatorientált, a statisztikai részletekben nem elvesző, a módszerek helyes alkalmazására azonban felkészítő, az eredmények értelmezésére nagy hangsúlyt fektető bevezetést is nyújt a kötet.

Bár a tárgyalt adatelemzési technikák nem csak a társadalomtudományi elemzéshez használhatóak, a társadalomtudományi jelleg két szempontból is fontos szerepet játszik a könyvben. A használt példák és adatbázisok társadalomtudományi adatokon alapulnak. Két adatbázis kerül alkalmazásra a könyvben. A legtöbbször a European Social Survey

[1] Sebők Miklós, Ring Orsolya, Máté Ákos (2021): Szövegbányászat és mesterséges intelligencia R-ben. Typotex Kiadó, Budapest.

(ESS) kilencedik hullámának adatait használják a példák. Az ESS egy kétévente, Európa számos országában lefolytatott kérdőíves kutatás, mely a válaszadók társadalmi hátterével, politikai viselkedésével és politikai, illetve társadalmi attitűdjeivel kapcsolatban tartalmaz adatokat. A kilencedik hullám 2018-ban került lekérdezésre 30 európai országban. A könyv az első fejezeteiben a teljes adatbázist használja ('ess'), az utolsó két fejezet azonban csak a magyarországi válaszadókat tartalmazó részadatbázisra ('hun') támaszkodik. Az ESS adatbázisai regisztráció után bárki számára elérhetőek a kutatás honlapján.[2] A másik adatbázis ('fb') a 2018-as magyarországi országgyűlési választások jelöltjeinek a kampány alatt közzétett összes Facebook-posztját tartalmazza. Azokat a jelölteket veszi figyelembe, akik vagy egyéni választókörzetben értek el 1%-nál jobb eredményt, vagy legalább 0,5%-os listás eredményt elérő párt listájának az első harminc helyének valamelyikén szerepeltek. Ez az adatbázis a könyv első felében kerül alkalmazásra.

A könyvben használt adatok és egyéb információk, így az egyes fejezetekben használt elemzések „scriptjei" elérhetőek és letölthetőek a https://github.com/benemarton89/r_adatelemzes oldalról. Az ESS adatbázisa azonban csak a kutatás honlapjáról tölthető le, ezért azt külön kell majd hozzáadnunk az adatainkhoz. Ennek mikéntjét az első fejezet végén részletezi a könyv: az itt bemutatott lépéseket mindenképpen végre kell hajtani ahhoz, hogy a könyv további fejezeteiben használni lehessen az ESS adatait. Az ötödik és hatodik fejezetben használt hun adatbázist szintén önállóan, az ötödik fejezet elején részletezett lépésekkel összhangban kell létrehozni.

A társadalomtudományi jelleg abban a szemléletben is érvényre jut, hogy a könyv nagy figyelmet fordít arra, hogy olyan eszközöket mutasson be, amelyek a minta reprezentativitását biztosító súlyozást is lehetővé teszik. Kérdőíves társadalomtudományi kutatásokban gyakori, hogy a minta csak ún. súlyok alkalmazásával tekinthető olyannak, mint ami reprezentálni tudja a populációt. A kérdőíves minták összeállításában gyakran érvényesül valamilyen torzítás, amit azonban a kérdőívek készítői a bekerülési valószínűségeket figyelembe vevő súlyok alkalmazásával korrigálnak. Ha például egy bizonyos jellemzőkkel bíró válaszadónak nagyobb esélye van a mintába kerülni, mint egy másik válaszadónak, akkor az adott súly alkalmazásával biztosítható, hogy előbbi válaszadó válaszai valamivel kisebb súllyal, utóbbié pedig nagyobb súllyal legyenek figyelembe véve. Ezek a súlyok a kérdőíves adatbázisokban külön változóként jelennek meg, és amennyiben alkalmazásra kerülnek, a súlyozással korrigált eredmények jelennek meg az elemző számára.

Ugyanakkor nem minden elemzési eljárás képes kezelni a súlyokat, hiszen nem társadalomtudományi területeken nem annyira jellemző a súlyok használata. Az R-ben való adatelemzést ismertető nemzetközi tankönyvek általában nem foglalkoznak a súlyozás problémájával, ez azonban azzal a következménnyel jár, hogy a súlyozott adatokkal dolgozni kívánó társadalomtudományi elemző az ezekben ismertetett eljárások egy jelentős részét

[2] https://www.europeansocialsurvey.org/

nem tudja alkalmazni. Emiatt ez a könyv az egyes elemzési eljárásoknál olyan megoldásokat mutat be, ahol van lehetőség az adatok súlyozására.

A társadalomtudományi képzésben részt vevő hallgatók a legtöbb esetben az SPSS statisztikai programcsomag használatán keresztül ismerkednek meg az adatelemzés alapjaival. Emiatt az értelmezést megkönnyítendő néhol történik utalás az SPSS programmal való hasonlóságokra és eltérésekre, ami segítheti az értelmezést azok számára, akik SPSS háttérrel kezdenek el foglalkozni a könyvvel. Ugyanakkor ezek csak az értelmezést segítő párhuzamok, nem szükséges az SPSS-t ismerni ahhoz, hogy valaki hatékonyan tudja használja e tankönyvet.

A könyv első két fejezete az R-program alapvető használatába vezeti be az olvasót. Az első fejezet a program felületét mutatja be részletesen, a második fejezet pedig az R-nyelv alapjaival ismerteti meg a felhasználót. Számos, a témában írt nemzetközi kötettel ellentétben ebben a fejezetben szándékosan csak az adatelemzéshez leginkább szükséges elemek kerülnek tárgyalásra, hiszen a kódokon keresztüli adatelemzésben járatlan olvasót könnyedén elriaszthatja, ha ezen a ponton olyan – a parancsírást hosszú távon valóban megkönnyítő – információkkal találkozik, mint a „for loop" használata vagy a funkcióírás miként je. A harmadik fejezet az adatbázis-menedzsmentbe vezeti be az olvasót, ami mindig is fontos részét képezte az adatelemzési ismereteknek, amióta azonban számtalan forrásból rengeteg eltérő formátumú adathalmaz érhető el és vált társadalomtudományi elemzések tárgyává, az adatok hatékony kezelése és alakítása még inkább nélkülözhetetlen tudássá vált. Az R-nek ráadásul előnye a rugalmas adatkezelési felépítés, ezért kifejezetten alkalmas különböző forrásokból származó adatok kezelésére. A negyedik fejezet az egyváltozós elemzésekbe nyújt bevezetést, a minta alacsony és magas mérési szintű változók mentén történő vizsgálatának miként jét tárgyalja. E fejezet nagy hangsúlyt fektet az elemzések ábrázolására is, amelyre némileg összetett, de rendkívül sokoldalú platformot biztosít az R ggplot2 csomagja. Az ötödik fejezet a kétváltozós elemzési technikákat tárgyalja, a kereszttábla-elemzés, a korrelációelemzés, a varianciaelemzés és a t-teszt kerül itt bemutatásra. A hatodik fejezet a többváltozós regressziómodellek világába vezeti be az olvasót. A különböző regressziós modellek (OLS és GLM) tárgyalásán túl számos modellezési szempont áttekintésre kerül, így a hierarchikus modellépítés vagy éppen az interakciós hatások kérdéskörét is bemutatja a fejezet.

Minden programnál, de a nyílt forráskódú programoknál különösen, számolni kell azzal, hogy a funkciók és csomagok nem „zártak" és véglegesek, hanem különböző frissítéseknek és javításoknak köszönhetően működésük és elérhetőségük az időben változhat. E könyv készülése közben is történtek változások, egyes funkciók a könyv írásakor nem pontosan úgy működtek, mint ahogy akkor, amikor a könyv szerkesztése zajlott. Éppen ezért könnyen előfordulhat, hogy az itt szereplő példák némelyike nem fut le, vagy éppen más eredményt mutat, mint ami a könyvben megjelenik. Ilyen esetben egy online keresés vagy éppen a funkcióhoz tartozó package hivatalos oldalán a frissítések áttekintése se-

gíthet megoldani a problémát. Előfordulhat, hogy a meglévő funkciót némileg máshogy kell alkalmazni, de az is megtörténhet, hogy másik funkciót kell keresni az adott feladat végrehajtására. Ugyanakkor kellő R-ben szerzett rutin birtokában, amelyhez reményeim szerint e könyv hozzásegít, az ilyen helyzetek kezelhetőek lehetnek.

Szintén potenciális hibaforrást jelent, ha egyes frissítések miatt a korábban egymásra épülő package-ek között valamilyen hézag keletkezik: a korábban működő funkciók nem futnak le, mert egy bizonyos package vagy annak a megfelelő verziója nincsen feltelepítve. Ilyen esetekben érdemes követni a hibaüzenet utasítását, vagy akár rákeresni a konkrét hibaüzenetre, hiszen online fórumokon számos technikai problémát megvitatnak a felhasználók. Végső esetben a program újratelepítése is meg tudja oldani a problémát.

Mindazonáltal az esetleges problémákat a könyvhöz tartozó GitHub-felületen is lehet jelezni, ahol a könyv szerzője igyekszik válaszolni az ilyen jellegű kérdésekre. Az „Issues" felületen bárki nyithat problémafület, de új probléma felvetése előtt érdemes leellenőrizni, hogy akár a nyitott, akár a zárt „issue"-k között nem vetődött-e már fel ugyanaz a kérdés. A GitHub-felület nyitóoldalát is érdemes előzetesen áttekinteni, mert a könyv esetleges hibáit, illetve a package-ekben és funkciókban azóta történt változásokat itt nyomon lehet követni.

Végezetül szeretném hálámat kifejezni azoknak, akik bármilyen módon hozzájárultak e könyv megszületéséhez. Először is nagy köszönettel tartozom a kötet szaklektorának, Papp Zsófiának, aki nagy alapossággal és szakmai hozzáértéssel vette kezelésbe a kéziratot, és számtalan hibára, pontosítási és javítási lehetőségre felhívta a figyelmemet. Szintén hálás vagyok Arató Krisztinának és Mándi Tibornak, akik lehetővé tették, hogy az ELTE ÁJK-n zajló politikatudományi mesterképzésben az R-programra épülő adatelemzési kurzust oktassak, mely nagyban motiválta, hogy e tankönyv megírásába fogjak. Szintén köszönettel tartozom a 2022 őszi és a 2023 őszi „Adatelemzés" mesterszakos kurzus hallgatóinak, valamint Burai Krisztinának, akik a könyv jelentős részét élesben tesztelték, és tapasztalataik megosztása mellett számos hibára felhívták a figyelmemet. Hálás vagyok Sebők Miklósnak is, akinek szakmai ajánlása sokat jelentett abban, hogy a könyv elnyerhette az MTA könyvkiadási támogatását, illetve a Ring Orsolyával és Máté Ákossal írt *Szövegbányászat és mesterséges intelligencia R-ben* könyvükkel inspirációt nyújtottak azzal kapcsolatban, hogy lehet értelme egy ilyen magyar nyelvű kötetnek. Köszönöm a Typotex Kiadó munkáját is a kézirattal, illetve Erő Zsuzsa korrektori munkáját. Végezetül köszönöm családomnak, Annának, Panninak és Ákosnak, hogy türelmükkel és szeretetükkel támogatták a munkámat.