DOKTORI (PhD) ÉRTEKEZÉS

HÁMORI GÁBOR

Kaposvári Egyetem

2014

KAPOSVÁRI EGYETEM

Gazdálkodás- és Szervezéstudományok Doktori Iskola

Doktori Iskola vezetője:

PROF. DR. Kerekes Sándor

MTA Doktora

Témavezető:

PROF. DR. SZÁZ JÁNOS

Egyetemi Tanár

PREDIKCIÓS CÉLÚ KLASSZIFIKÁLÓ STATISZTIKAI MODELLEK GYAKORLATI KÉRDÉSEI

Készítette:

Hámori Gábor

KAPOSVÁR

2014

DOI: 10.17166/KE.2015.007

Tartalomjegyzék

Táblázatok jegyzéke

Ábrák jegyzéke

A disszertáció célja, hipotézisek

Napjainkra a modern adatgyűjtési és adattárolási technológiáknak köszönhetően a jelentős tevékenységi volumennel jellemezhető gazdasági szereplők elektronikus rendszereiben óriási adatvagyonok halmozódtak fel. A továbbiakban már nem az a kérdés, hogy honnan szerezzünk adatokat, hanem az, hogy mit kezdjünk a rendelkezésre álló hatalmas adattömeggel.

A nagyméretű adatbázisokból történő, stratégiai döntések megalapozását szolgáló nem triviális információk kinyerésének korszerű eszköze a „data mining" (adatbányászat), vagy „big data" fogalmakkal leírt technológiák. Ezen technológiák kialakulása és fejlődése napjainkban is tart, és többek között olyan tudományágak szintéziseként írhatjuk le őket, mint azinformáció technológia, adatbázis tervezés, adat vizualizáció, gépi tanulás, sokváltozós statisztikai analízis és modellezés. A gyakorlati adatmodellezés talán egyik leggyakoribb feladata bizonyos mennyiségek, illetve bizonyos jövőbeli események bekövetkezésének előrejelzése, predikciója. Gondoljunk csak egy telekommunikációs vállalatra, aki szeretné előrejelezni, hogy ügyfelei egy meghatározott időszakban mekkora valószínűséggel fognak lemorzsolódni, azaz a legtöbb esetben egy másik szolgáltatóhoz átpártolni. Vagy jellemző példa a kereskedelmi bankok, ügyfelek nemfizetésére vonatkozó előrejelzése, ismertebb nevén a hitelscoring. Sok esetben az adóhatóságok is prediktív modelleket használnak a vizsgálatba bekerülő adóalanyok kiválasztására. A példákat hosszasan lehetne sorolni.

A predikciós modellezés hagyományos és közkedvelt eszköze a sokváltozós regressziós technikák alkalmazása. A regressziós algoritmusok többsége könnyen hozzáférhető és alkalmazható, hiszen gyakorlatilag minden statisztikai vagy adatbányász szoftverben megtalálható egy vagy több típusuk, beleértve a nyílt forráskódú programcsomagokat is. Továbbmenve az előrejelzési célú (predikciós) regressziós modellezés gyakorlatához kapcsolható technikai részletkérdések külön-külön is nagyon jól feltérképezett területnek tekinthetők, az egyes problémákra fokuszálva nagyon sok mérvadó tanulmány született az elmúlt évtizedekben. A szerzők általában a kérdések mögött meghúzódó matematikai háttér igényes bemutatásával és igénybevételével izoláltan, a modellkészítés folyamatából kiragadva, elemzik az egyes részterületeket, problémákat. Az alkalmazott statisztikai modellezés, a szélesebb értelemben vett adatbányászati megoldások rohamos és iparszerű

elterjedésével azonban megjelent egy új igény az egyes, a modellfejlesztés során felmerülő kérdéseket a modellezés folyamatában vizsgáló és elemző megközelítés iránt.

A disszertáció a gyakorlati alkalmazásokban egyik legelterjedtebb, predikciós célú logisztikus regressziós modellt a vizsgálódás középpontjába helyezve, tesz kísérletet ennek a statisztika területén újszerű szemléletmódnak az érvényesítésére. A tanulmány segítségével átfogó képet kapunk a teljes modellezési folyamatról, az egyes lépéseknél jelentkező tipikus döntési helyzetekről, és az azokra adható válaszokról, legjobb gyakorlatokról. A hivatkozások segítségével az egyes részletkérdések iránt mélyebb ismeretre vágyók is útmutatást kapnak. A disszertáció így mind a gyakorló, mind az elméleti szakemberek érdeklődésére is igényt tarthat.

A tanulmány egyik fő célja tehát az, hogy a bináris klasszifikációra alkalmas dichotóm logisztikus regresszióra épülő előrejelző modellezés gyakorlatát és az ezzel kapcsolatos célkitűzéseket, módszertani problémákat lépésről lépésre bemutassa. Habár a mondanivaló a logisztikus regressziós modell köré épül, a leírtak gyakran érvényesek más regressziós (leggyakrabban lineáris) modellek esetében is, amit a szövegben megfelelő hivatkozással vagy utalással meg is jelölünk. A dolgozat elkészítésénél alapvető fontosságú volt a leírtak gyakorlati alkalmazhatósága és az elméleti, módszertani igényesség. Teljes körűségre a téma összetett és szerteágazó mivolta miatt nem törekedhettünk, ezért igyekeztünk a modellezési folyamat során leggyakrabban előforduló, legjellemzőbb helyzeteket modellépítés során előfordulásuk sorrendjében elemezni és értékelni. Az egyes problémáknál a megfelelő szakirodalom áttekintésével vagy meghivatkozásával próbáljuk az érdeklődő olvasót segíteni abban, hogy további ismereteket szerezhessen az adott módszertani kérdéssel kapcsolatban.

A predikciós célú klasszifikáló modellek gyakorlati alkalmazása esetében kitüntetett jelentőségű az elkészült modell illeszkedése, klasszifikáló ereje. A disszertációban ezért megvizsgálunk három olyan hipotézist, melyek segítségével a logisztikus regressziós modell illeszkedése jelentősen javítható.

- ***Hipotézis 1***: A modellezési adatbázisban a szélsőséges, kiugró (outlier) értékek megfelelő kezelésével a modell illeszkedése javítható.
- ***Hipotézis 2***: A folytonos változók kategorizálása még abban az esetben is növelheti a modell prediktív erejét, amennyiben célváltozóval az eredeti folytonos változó monoton kapcsolatban van

- ***Hipotézis 3***: Ismeretes, hogy a folytonos változók kategorizálása során az egyes kategóriákat meghatározó kategóriahatárok számának és elhelyezkedésének megválasztása befolyásolja a változó prediktív erejét a modellben. A CHAID-algoritmus, disszertációban ismertetésre kerülő, részalgoritmusa alkalmas az illeszkedés maximálása szempontjából optimális kategóriahatárok kialakítására.

Intuitív megfontolások alapján a hipotézisekben megfogalmazott illeszkedésjavító módszerek várhatóan más regressziós modell (pl. lineáris modellek) típusok esetében is növelhetik a modell klasszifikáló képességét. Ennek igazolása, túlmutatva jelen disszertáció keretein, jövőbeli kutatások tárgya lehet.

A disszertáció felépítése

A *bevezetés* során áttekintjük és összevetjük a statisztikai és a szakértői predikciókat tulajdonságaik és alkalmazhatóságuk jellemzői mentén. A statisztikai predikciók egyik legfőbb részterületén, a klasszifikációk terén vizsgálatunkat leszűkítjük a gyakorlati alkalmazásokban leggyakrabban alkalmazott klasszifikáló előrejelzésre alkalmas statisztikai eszközre, a logisztikus regressziós modellre Ennek megfelelően a disszertáció első része a *dichotóm logisztikus regresszió*, mint statisztikai modell bemutatásával és legfontosabb részletkérdéseivel foglalkozik. A modell definiálását követően bemutatásra kerül a paraméterbecslés maximum likelihood eljárása, és a modell teljesítményét értékelni hivatott illeszkedésvizsgálatok. A második rész már magáról a modellezési folyamat első lépéséről szól. *A modellezési adatbázis kialakításának szempontjai* fejezet a modellezés alapjául szolgáló adatállomány sarkalatos kérdéseivel foglalkozik. Az adatmátrix szerkezeti arányaiból eredő sajátosságokon túlmenően foglalkozunk a kis és kiegyensúlyozatlan minták, valamint az adatbázis tesztelés célú particionálásának kérdésével is. A következő részben a gyakorlatban szinte „törvényszerűen" előforduló *hiányzó értékek*problematikáját elemezzük. A probléma strukturálása mellett sor kerül az egyes megoldási módok ismertetésére és értékelésére is. Ezt követően kerül sor az adatbázis modellezést torzító, szélsőséges értékeinek az úgynevezett *outlier értékek és kezelésük* tárgyalására. A klasszifikációs erő fokozása céljából külön fejezetben foglalkozunk a lehetséges magyarázó

változók előzetes transzformációjával. A modellezéssel kapcsolatban sok esetben elvárás, hogy az elkészült modell segítségével egy adott egyed esetében konkrét valószínűségbecslést tudjunk adni. Ennek részletkérdéseivel foglalkozik a *valószínűségbecslés modellel* című fejezet. Az utolsó fejezetben térünk ki arra az esetre, amikor kitűzött feladatunk nem pusztán egy nagyon jól illeszkedő modell megalkotása, hanem vizsgált jelenséget befolyásoló faktorok, és azok kapcsolatrendszerének feltérképezése és megértése is a célunk. Az *exploratív modellezés*t leíró fejezetben kitérünk a logisztikus regresszió paramétereinek értelmezésére, megvizsgáljuk, hogy mennyiben kell eltérnünk a maximális prediktív erővel bíró modell megépítéshez vezető folyamathoz képest. A leírtak szemléltetésére a fejezetet egy részletes *esettanulmánnyal* zárjuk.

I.Bevezetés

I/1.Statisztikai versus szakértői előrejelzések

Egy internetes könyváruházban keresgélve nem szokatlan, hogy az oldal automatikusan felkínál számunkra bizonyos könyveket, melyek jó eséllyel számíthatnak az érdeklődésünkre. Telefonszolgáltatónktól sms-t kapunk, melyben célirányosan javasolnak új terméket számunkra. Banki hitelért folyamodva, kérelmünket ma már pár perc alatt automatikusan elbírálja a hitelintézet. A felsorolt példákban az a közös, hogy minden esetben számítás intenzív prediktív módszerek segítségével történtek előrejelzések várható viselkedésünkre vonatkozóan.

A számítástechnika területén megfigyelhető technológiai forradalom az elmúlt évtizedekben megfigyelhető robbanásszerű fejlődése lehetővé tette, hogy másodpercek alatt több millió műveletet hajthassunk végre személyi számítógépek segítségével. Korábban a bonyolult matematikai-statisztikai algoritmusok (továbbiakban: algoritmusok) alkalmazása a gyakorlatban, inkább csak az akadémiai kutatások körében volt tetten érhető végrehajtásuk jelentős időigénye és a korlátozott számítástechnikai kapacitás miatt. Mára azonban a helyzet gyökeresen megváltozott. Külön iparág jött létre azzal a céllal, hogy az algoritmusokat felhasználók részére elérhetővé tegye. Rendre váltak elérhetővé a különböző célszoftverek, melyek segítségével korábban akár hetekig tartó elemzéseket pár perc alatt végrehajthatjuk. Hála a nyílt forráskódú szoftvereknek ma már az algoritmusok jelentős része bárki számára hozzáférhető.

A számítási lehetőségek gyors ütemű bővülése és a könnyű hozzáférés természetesen további lökést adott az alapkutatásoknak ezen belül is az algoritmusfejlesztésnek. Szakmai tudományos műhelyekben rendszeresen lát napvilágot egy-egy új eljárás vagy algoritmus, melyek közül a legéletképesebbek hamar meghonosodnak a gyakorlatban is.

Az alkalmazás területén elsőként a vállalati/üzleti szféra ismerte fel az új technológiában rejlő lehetőségeket. Az olyan iparágak, ahol jellemző a tömegigényeket kielégítő termelés illetve szolgáltatás, az üzleti folyamatok statisztikai elemzése és modellezése magától értetődő. A tömegszerű döntéseket igénylő helyzetekben a korszerű nagyvállalatok korán felismerték, hogy algoritmusok eredményeképp előálló prediktív döntési szabályok/formulák segítségével gyorsan, és ami még talán ennél is fontosabb, olcsón

tudnak dönteni. Ellentétben a szakértők általi döntésekkel ezek az algoritmus alapú döntési szabályok mentesek a fáradtságtól, kognitív torzítástól, nélkülözik a szubjektív elemeket és időben koherensek, azaz ugyanazt a döntési szituációt két különböző időpontban ugyanúgy értékelik. Erre jó példa az a kutatás (P.J.Hoffman,P.Slovic,L.G.Rorer,1968) melynek során tapasztalt radiológusokat kértek meg arra, hogy mellkasröntgen felvételeket értékeljenek a „normális" és „abnormális" kategóriák mentén két különböző időpontban, úgy hogy a kísérletben résztvevők nem tudták, hogy másodszorra ugyanazokat a felvételeket látják. Az esetek 20 %-ban az értékelések önellentmondóak voltak. Hasonló mértékű következetlenséget figyeltek meg abban a kutatásban, melynek során 101 könyvvizsgálót kértek fel arra, hogy értékeljék a vállalati belső ellenőrzések megbízhatóságát (P.R.Brown 1983).

Azokban a szituációkban, amikor a vizsgált probléma komplexitása, „mérete" jelentős, a döntési helyzet összetettsége következtében az algoritmusok olyan összefüggések feltárásárais képesek, melyet az emberi elme már nem tud megragadni. Gondoljunk például egy félmillió ügyféllel, és ennek megfelelően rengeteg leíró adattal rendelkező hitelintézetre, ahol például arra keresik a választ, hogy: Mitől és hogyan függött, hogy valaki visszafizette-e a kapott hitelt vagy sem?- vagy: Lehetséges-e valószínűségi becslést adni egy új ügyfél esetében a nemfizetés valószínűségére, és ennek alapján dönteni, hogy ki kapjon hitelt?

Megfelelő statisztikai elemzések segítségével jellemzően feltárhatók mindazok a tényezők, melyek hatást gyakorolnak a fizetési hajlandóságra, valamint alkalmas algoritmusok segítségével előállítható olyan formula, melynek segítségével meg lehet becsülni egy konkrét ügyfél esetében a nemfizetési valószínűségét. A feltárt összefüggések alapján olyan üzleti döntési szabályok hozhatók létre, melyek egy célfüggvény mentén optimális folyamatot eredményeznek és azok az alkalmazás során automatikusan végre hajtásra kerülnek[1].

A fenti döntési helyzet összetettsége okán talán nem meglepő, hogy ezeken a területeken az emberi (szakértői) döntés/becslés rosszabbul teljesít, mint az algoritmus-alapú szabály. A meglepő azonban az, hogy sok esetben ugyanez figyelhető meg „kismintás" esetben is, azaz ahol a probléma jóval kevesebb adattal írható le. Paul Meehl 1954-ben megjelent *Clinical*

[1] A fenti két kérdéshez kapcsolódóan jó példa a nagybankok által használt automatikus hitel elbírálási ún. scoring rendszerek alkalmazása. Ezek a rendszerek a megadott adatok alapján teljesen automatikusan, emberi tényező teljes kizárásával, másodpercek alatt elvégzik az hitelkérelmező ügyfél nemfizetési valószínűség becslését, és ennek alapján tesznek javaslatot az szerződés megkötésére, illetve magas nemfizetési valószínűség esetén, a kérelem elutasítására.

*vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*című munkájában 20 olyan kutatás eredményeit foglalta össze, amelyek azt vizsgálták, hogy a képzett szakemberek szubjektív értékítéletén alapuló klinikai előrejelzései pontosabbak-e, mint egy adott szabályból származtatható statisztikai előrejelzés. A kutatás eredménye szerint a statisztikai előrejelzések jellemzően sokkal jobban teljesítettek, mint a szakértők általi becslések. Még meglepőbb a 2002-ben közgazdasági Nobel díjjal kitüntetett Daniel Kahneman eredménye a témában, aki a 2011-ben megjelent *Gyors és lassú gondolkodás* című művében a következőket írja (258 o.) ”*A klinikai és statisztikai előrejelzések összehasonlításáról beszámoló kutatások száma mintegy 200-ra nőtt, de az algoritmusok és az emberi előrejelzések közötti verseny eredményei nem változtak . A kutatások közel 60%-ban az algoritmusokat jelentősen pontosabbnak találták. A többi összehasonlítás döntetlen eredményt hozott a pontosság szempontjából, de a holtverseny valójában a statisztikai szabályok győzelmét jelentette, mivel ezek általában sokkal olcsóbbak, mint a szakértői ítéletek. A kutatások egyetlen meggyőző kivételről sem számoltak be.”* További váratlan eredménye a kutatásoknak, hogy a statisztikai predikciók fölénye még a magyarázó faktorok egyszerű lineáris kombinációjaként előálló szabályok esetén is megfigyelhetők.

A statisztikai előrejelzések alapjául szolgáló modellek az előrejelzésen túl arra is alkalmasak, hogy feltárják az előrejelzés magyarázó faktorait, azok összefüggéseit egymással és az előrejelezni kívánt mennyiséggel. Ezáltal nemcsak jól működő „kristálygömbként” hasznosíthatók az előrejelzés során, hanem általuk tudásra, a modellezendő jelenség belső mechanizmusának, összefüggés rendszerének megértésére is szert tehetünk. Leegyszerűsítve azt is mondhatnánk, hogy egy megfelelő adatbázison egy jó statisztikai modell pár perc alatt képes „megtanulni” mindazt, amihez a terület szakértőjének esetleg több évtizedre volt szüksége.

Az eddigiekből talán arra következtethetünk, hogy az előrejelzési szituációkban teljesen mellőzhető az emberi bölcsesség, a szakértői tudás. „Szerencsére” a statisztikai alapú algoritmikus előrejelzések sem mentesek bizonyos hátrányoktól. Az egyik ilyen a Paul Meehl által „törött láb” jelenségnek (broken leg phenomenon) nevezett helyzet. Meehl gondolatkísérletében feltette, hogy rendelkezünk egy statisztikai algoritmussal, mely a korábbi tapasztalatok alapján nagy biztonsággal képes előrejelezni, hogy egy bizonyos professzor szerda este moziba fog menni. Az algoritmus remekül működik, mindaddig, míg egy keddi napon a professzor váratlanul el nem töri a lábát és így szerdán nem tud moziba menni. Tehát az algoritmusok rosszul teljesítenek olyan helyzetekben, amikor bekövetkezik

egy korábban nem megfigyelt alacsony valószínűségű ritka esemény, melynek a kimenetre gyakorolt hatása jelentős[2]. A másik probléma az algoritmusfejlesztés kiindulását jelentő adatszerkezettel kapcsolatos. Az előrejelző modellek csak az adatok által reprezentált világban képesek jó eredményeket szolgáltatni. Ha valamilyen oknál fogva az adatbázisból kimaradnak lényeges, a prediktálni kívánt mennyiség alakulását erősen befolyásoló változók, az előrejelzés minősége romlani fog. A gyakorlati alkalmazásnál további fontos szempont az adatok megbízhatóságának, minőségének a kérdése. A felhasznált adatok minősége meghatározó a modell előrejelzési képessége szempontjából. Ha nem megfelelő adatokat használunk, az eredmény sem lesz használható. Ezt a jelenséget fejezi ki az úgynevezett GIGO elv ("Garbige In Garbige Out").

Gyorsan változó, alacsony strukturáltsági szintű adathiányos döntési helyzetekben tehát az emberi bölcsesség továbbra is nélkülözhetetlennek tűnik. A statisztikai előrejelzések elsőbbsége viszont megkérdőjelezhetetlen azokban a környezetekben, ahol jellemző a döntések tömegszerűsége illetve az információk jó minőségben, strukturáltan és széleskörűen állnak rendelkezésre. Manapság tipikusan ilyen környezet a modern nagyvállalat[3]. További példa lehet a nagy állami rendszerek, mint például az egészségügy, a nyugdíjrendszer vagy akár az adóhatóság. A továbbiakban röviden áttekintjük, hogy melyek azok a statisztikai előrejelzési feladatok, amelyek az említett területeken leggyakrabban fordulnak elő.

I/2. Statisztikai előrejelzés és regresszió

A leggyakrabban előforduló előrejelzési feladatokat statisztikai szemszögből két részre oszthatjuk. Az egyik, amikor az előrejelezni kívánt érték mennyiségi skálán értelmezhető. Várható adóbevételek, vagy adott termék értékesítésének predikciója tipikusan ilyenek. Az ilyen típusú előrejelzések népszerű technikája a többváltozós lineáris regresszió. A másik előrejelzési helyzet, amikor az előrejelzés valamilyen csoportba történő tartozásra vonatkozik, azaz a célváltozó kategória kimenetelű. Ilyenkor beszélünk klasszifikációról. A gyakorlatban előforduló klasszifikációs feladatokat hosszasan lehetne sorolni. Gazdasági súlyát tekintve egyik legfontosabb alkalmazás a nagyvállalati, elsősorban banki, ügyfelek

[2] Gazdasági rövidtávú előrejelzésekben ilyen atipikus esemény a trendforduló vagy a válsághelyzet
[3] Legjellemzőbb iparágak: Bank/biztosítás, Gyógyszeripar, Telekommunikáció, Kereskedelem és tömegtermelés

nemfizetésének előrejelzése (scoring). Ilyenkor a klasszifikáció a „fizet” és „nem fizet” csoportokba történő tartozás előrejelzésére vonatkozik. Másik, elsősorban a telekommunikációs vállalatok esetében, kiemelt jelentőségű előrejelzési feladat, az ügyfelek várható lemorzsolódásának (churn) előrejelzése. Az ügyfelek jövőbeli várható viselkedése alapján jellemzően itt is két kategóriába, a „marad” és „lemorzsolódik” kategóriákba történő besorolás a klasszifikációs feladat.A statisztikai klasszifikációt széleskörűen alkalmazzák csalások felderítésénél is. Az adóhatóságok például statisztikai algoritmusokat használnak annak megállapítására, hogy egy adott „ügyfél” milyen valószínűséggel lesz adóelkerülő, illetve rendben fizető. Egészségügyi alkalmazásokban és kutatásokban gyakori kérdés, hogy egy adott betegség, vagy állapot milyen eséllyel következik be egy konkrét beteg esetében.

A felsorolt példák mutatják, hogy a bináris klasszifikáció milyen nagy súllyal szerepel napjainkban előforduló előrejelzési feladatok között. A felsorolt előrejelzési szituációkban a modellezés célja általában kettős. Egyrészt kiemelten fontos cél a jó klasszifikációs teljesítmény, a jó „előrejelzési képesség”. Ez különösen fontos a profitorientált vállalatok esetében. Másrészt a modellezés segítségével szeretnénk megérteni, hogy mik azok a tényezők, amelyek egy adott esemény bekövetkezéséhez, vagy elkerüléséhez vezetnek. A kitűzött célok elérésére számos statisztikai eljárás alkalmazható.

A bináris klasszifikációra alkalmas statisztikai algoritmusok közül a gyakorlati alkalmazásokban legnépszerűbb, leggyakrabban alkalmazott hagyományos technika a dichotóm logisztikus regresszió. Közkedveltségét több tényezőnek köszönheti. Egyrészt kevés megkötést alkalmaz a felhasználásra kerülő adatok vonatkozásában, így robusztus módszernek tekinthető. Megfelelő előrejelzési teljesítmény párosul azzal, hogy a modell jól interpretálható, megfelelő feltételek megléte esetén az esemény bekövetkezését befolyásoló tényezők, azok előrejelzésben betöltött súlya a modell segítségével azonosítható. Elterjedésében nem csekély szerepe van annak, hogy gyakorlatilag minden piacon elterjedt statisztikai/adatbányászati szoftverben elérhető. Ehhez kapcsolódik, hogy a felsőoktatásban a haladó, többváltozós statisztika témakörben törzsanyagként kerül oktatásra hosszú évek óta. A módszer iránti érdeklődést az is jelzi, hogy nagyon jól kutatott terület, rengeteg publikáció született a logisztikus regresszió különböző technikai részletkérdéseivel kapcsolatban az elmúlt évtizedekben. A következő fejezetben áttekintjük a logisztikus regressziós modell főbb aspektusait.

II.Adichotóm logisztikus regresszió

II/1.A dichotóm logisztikus regresszió fogalma

A logisztikus regresszió olyan nemlineáris klasszifikációs eljárás, melynek segítségével, olyan kérdésekre kaphatunk választ, mint például: Mitől és hogyan függ, hogy valaki kap-e szívinfarktust, avagy sem; visszafizeti-e egy vállalat a hitelét, vagy csődöt jelent; folytatja-e tanulmányait egy érettségizett diák, vagy sem; fiú lesz-e a születendő gyermek, vagy leány.
A felsorolt kérdésekben közös, hogy a kimenet mindig két kategória valamelyikébe tartozik (igen-nem, fiú-leány, stb). Az eljárás alkalmas arra, hogy segítségével megragadjuk a kategóriákba esés szempontjából releváns magyarázó változók körét, azok hatását értelmezzük (exploratív szakasz), majd a megfigyelésekhez megfelelően illeszkedő logisztikus regresszió modell segítségével az új eseteknél valószínűségi kijelentést tegyünk a két kategóriába sorolódásra vonatkozóan (prediktív szakasz).
A logisztikus regresszió számos előnnyel rendelkezik más klasszifikációs eljárásokhoz képest:

- Nem tesz semmilyen megkötést a magyarázó változók eloszlásával kapcsolatban[4]
- Folytonos változókon kívül a logisztikus regressziós modellbe beépíthetők kategoriális mérési szintű változók is.
- A logisztikus regressziós függvény értékei adott feltételek mellett valószínűségeknek tekinthetők, amennyiben a logisztikus regresszió dichotóm függő változója a 0 és az 1 értékkel jelölt, a regressziós érték az 1-gyel jelzett kategóriába esés valószínűségét adja meg.
- Az eredmények jól interpretálhatók.

[4] A szintén közkedvelt klasszifikációs eljárás, a diszkriminancia analízis esetében például követelmény a magyarázó változók együttes eloszlására vonatkozó többdimenziós normalitás

II/2. Alapfeltevések

A logisztikus regressziós modell megépítésekor a következő alapfeltételezésekkel élünk:
A függő változónak dichotómnakkell lennie, mely két értéket vehet fel: 1-est $P(1)$és 0-át$P(0) = 1 - P(1)$valószínűséggel.
Egy megfigyeléshez csak egy kimenet tartozhat, de minden kimenethez kell, hogy tartozzon megfigyelés, más szavakkal nem fordulhat elő olyan eset, hogy egyszerre két kategóriába (0, 1) tartozzon, illetve minden függő változó értékhez kell, hogy tartozzanak a független változóknak értékei.
Tekintettel arra, hogy a logisztikus regresszió paraméterbecsléséhez a későbbiekben ismertetendő maximum-likelihood becslés szolgál eszközül, a módszer nagymintás kedvező tulajdonságai miatt kis minták alkalmazása kerülendő.

II/3. A logit

A logisztikus regressziós modell központi eleme az átviteli függvény, az ún.logit transzformáció[5]

$$z \to \log \frac{z}{1-z}, \tag{2.1}$$

ahol feltételezzük, hogy a feltételes várható érték logit transzformációja lesz a magyarázó változó értékeinek lineáris függvénye. Az egyszerűség kedvéért jelöljük$P(\boldsymbol{x})$-szel az $E(Y|\boldsymbol{X} = \boldsymbol{x})$ feltételes várható értéket, megjegyezve, hogy ez egyszersmind az$Y = 1$ esemény feltételes valószínűsége, azaz

$$P(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x}) = Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \tag{2.2}$$

Ekkor a fenti modell szerint a feltételes várható érték logit transzformációjalog$\frac{\mathrm{P(x)}}{1-\mathrm{P(x)}}$a magyarázó változó értékeinek (konstans tagot is tartalmazó) lineáris függvénye:

[5] logit link function

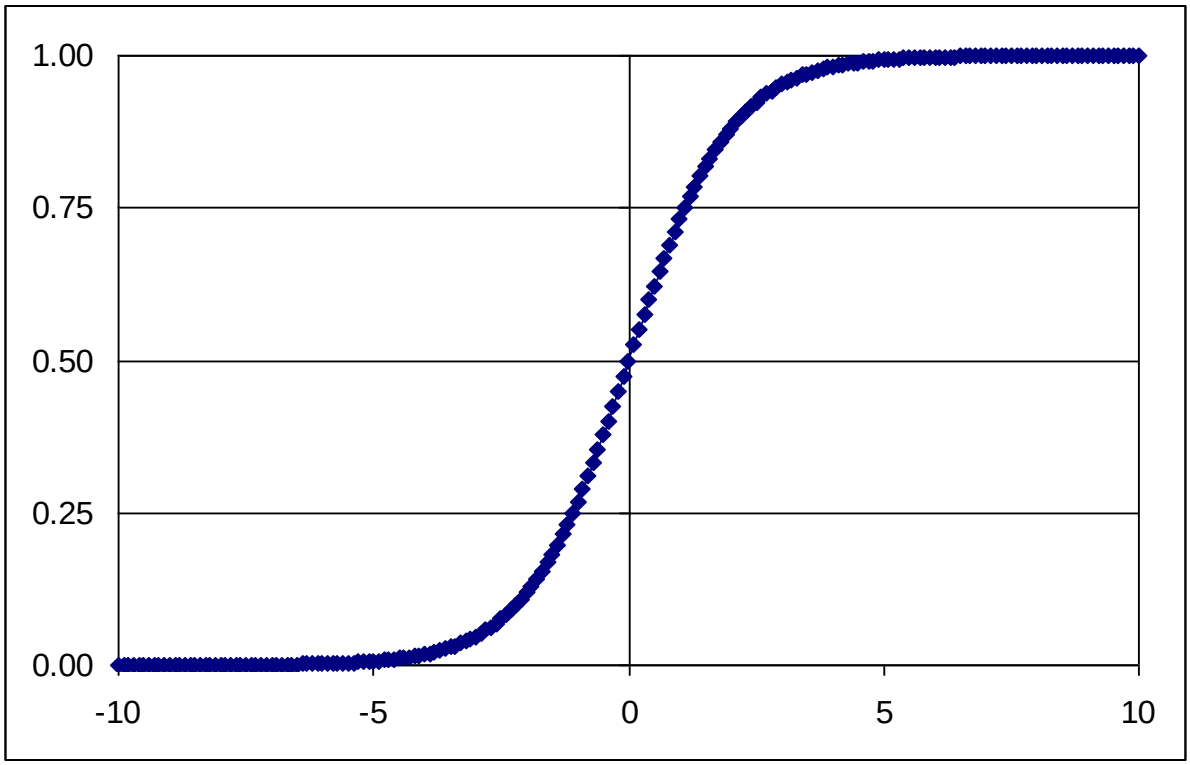
$$g(\boldsymbol{x}) = \log\frac{P(\boldsymbol{x})}{1-P(\boldsymbol{x})} = \beta \cdot x = \beta_0 + \sum_{j=1}^{k} \beta_j x_j \qquad (2.3)$$

A g(**x**) logit kedvező tulajdonsága, hogy paramétereinek folytonos függvénye és -∞-től +∞-ig veszi fel értékeit. A (3) formulából következően a feltételes várható értéket a magyarázó változó értékeinek függvényeként az alábbi alakban adhatjuk meg:

$$P(\boldsymbol{x}) = \frac{e^{\beta \cdot x}}{1 + e^{\beta \cdot x}}\,. \qquad (2.4)$$

A függvény alakja az alábbi ábrán tanulmányozható.

1. ábra Az exp(*y*) /(1+exp(*y*)) függvény képe



A logit transzformáció alkalmazásának célszerűsége és értelme könnyen magyarázható. A (3) képlettel adott függvény összetett függvény: a valószínűséghányados (esély), az úgynevezett

$$\boldsymbol{odds_x} = \frac{P(\boldsymbol{x})}{1 - P(\boldsymbol{x})} \tag{2.5}$$

logaritmusa. A valószínűséghányados – az esemény bekövetkezési és nem bekövetkezési valószínűségének hányadosa – gyakran használt mutató, a [0,1]-beli valószínűségeket az $\mathbf{R}^+$ -ba képezi. Az **odds** mutató kedvező tulajdonsága, hogy értelmezése könnyen követhető szabályok szerint lehetséges:

- olyan eseményre, ami inkább bekövetkezik mint nem, az esélyhányados 1 feletti;
- nagy valószínűségek mellett tetszőlegesen nagy értékeket is felvehet;
- 0 valószínűségű eseményekre az értéke nulla.

Hátránya, hogy csak a pozitív számok tartoznak bele az értékkészletébe. Ezen 'segít' a második függvény, a logaritmus. A logaritmusnak további haszna, hogy emellett a sokat használt likelihood függvény könnyen számolható lesz.

II/4. Paraméterbecslés

A logisztikus regresszió paramétereinek becslése maximum likelihood módszerrel történik. Ennek menete a következő (Hajdu,2003):
Legyen $i= 1,2,...,n$ független megfigyelésünk az y={1,0} dummy jellegű eredményváltozó $y_i=y|x_i$értékének alakulására a magyarázó változók $x_{i1}, x_{i2},...x_{ik}$ kovariánsa mellett. (A kovariánsok között azonosak is lehetnek. Egy adott kovariáns előfordulási gyakorisága legyen n_x.)

A logisztikus regresszió szerint az x_i kovariáns melletti y kimenet valószínűsége:

$$\Pr(y = y_i) = \frac{e^{y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}} \tag{2.6}$$

Ahol a β regressziós paramétereket becsülnünk kell. A maximum likelihood módszer szerint a paraméterek olyan b becsült értékeit keressük, amelyekre a minta bekövetkezésének likelihoodja (valószínűsége) (L) maximális:

$$L = \prod_{i=1}^{n} \Pr(y = y_i|\boldsymbol{b}) = \prod_{i=1}^{n} P_{ib}^{y_i} (1 - P_{ib})^{1-y_i} = \prod_{i=1}^{n} \frac{e^{y_i \boldsymbol{bx_i}}}{1 + e^{\boldsymbol{bx_i}}} \rightarrow max \qquad (2.7)$$

ahol P_{ib} az „1” kimenet valószínűségének az i mintaelem x_i kovariánsa mellett a b paraméterek felhasználásával becsült értéke.

Mivel a kovariánsok között azonosak is lehetnek, a likelihood felírható az alábbi súlyozott formában is:

$$L = \prod_{minden\ \boldsymbol{x}\ kovari ánsra} \mathrm{P}_{xb}{}^{f_x} (1 - \mathrm{P}_{xb})^{n_x - f_x} \qquad (2.8)$$

Ahol f_x az **x** kovariáns mellett bekövetkezett „1” kimenetek megfigyelt gyakorisága a mintában.Az L likelihood függvény maximalizálása helyett gyakran az lnL loglikelihood maximumhelyét szokás keresni a számítások egyszerűsítése végett. (Mivel monoton transzformáció nem változtatja a függvény szélsőérték helyét, ezt megtehetjük.)Ha az lnL loglikelihood függvény **b** szerinti deriváltjait vesszük, és ezeket nullával egyenlővé téve b-re megoldjuk, megkapjuk az úgynevezett maximum likelihood megoldást, ami definiálja azt a paramétervektort, amely tartalmazza azokat az együtthatókat, amelyekre az adott minta a legvalószínűbb. Ez a becslőfüggvény azonban jelen esetben nem írható fel zárt alakban, de nemlineáris egyenletmegoldás útján a keresett paraméterek meghatározhatók. Az ilyen feladatokra a statisztikai programcsomagok jól kidolgozott eljárásokat ajánlanak.

Az lnL loglikelihood maximálását a jóval általánosabb feladatosztály megoldására is alkalmas iteratív módon újrasúlyozott Gauss-Newton nemlineáris legkisebb négyzetek módszerrel is elvégezhetjük a következő módon:

Mivel a modell szerint az „1” kimenetelek várható száma az x kovariáns esetén

$$E(f_x) = n_x P_x \qquad (2.9)$$

az „1” kimenetelek számának varianciája pedig

$$Var(f_x) = n_x P_x (1 - P_x) \qquad (2.10)$$

Ezért a paraméterek becsléseit a

$$\sum_{minden\ \boldsymbol{x}\ kovari\ ánsra} \frac{(f_x - n_x P_{xb})^2}{n_x P_{xb}(1 - P_{xb})} \qquad (2.11)$$

súlyozott négyzetösszeg minimálásával kaphatjuk. Mivel a súlyokban (a nevezőben) is szerepelnek a modellből becsült értékek, ezért a súlyok újraszámítódnak minden iterációs lépésben. (A paraméterekre induló megoldást kell adni, amit lépésről lépésre javítunk a végső becslésig.)A módszer iránt érdeklődők például (Alfonso, Lindsey, Winnie, 2012) tanulmányából tájékozódhatnak részletesebben.

II/5. Paraméterek szignifikanciájának tesztelése

A becsült paraméterek együttes szignifikanciájának tesztelésére, a többváltozós lineáris regresszionál használatos globális F-próba mintájára, az ún. G-statisztika, másnéven maximum likelihood-arány (LR-Likelihood Ratio) teszt alkalmas. A statisztika alakja a következő:

$$G = \chi^2 = -2 \ln \left(\frac{L_{null}}{L_k} \right) \qquad (2.12)$$

Ahol a jobb oldali kifejezés számlálójában a csak konstanst tartalmazó úgynevezett nullmodell likelihood-ja, míg a nevezőben a k különböző prediktor tartalmazó modell likelihood-ja. A statisztika k darab magyarázó változó esetén k-szabadságfokú khi-négyzet eloszlást követ. A teszt nullhipotézise:

$$H_0: \qquad \beta_1 = \beta_2 = \beta_3 = \ldots\ldots = \beta_k = 0 \qquad (2.13)$$

Amennyiben adott szignifikancia-szint mellett a G–statisztika értéke magasabb, mint az ehhez a szinthez tartozó kritikusérték, a nullhipotézist elutasítjuk, azaz legalább egy magyarázó változó sokasági paramétere különbözik nullától az adott szignifikancia-szinten. Az egyedi paraméterek szignifikanciájának a tesztelésére, a lineáris regressziónál is megszokott, t-statisztika alkalmas. A statisztika a becsült paraméter (β_j) és standard hibájának hányadosaként áll elő, mely nagy minták esetén standard normális eloszlást követ. A teszt nullhipotézise:

$$H_0: \qquad \beta_j = 0. \tag{2.14}$$

A próba mind egyoldali, mind kétoldali módon végrehajtható, adott szignifikancia szinten.

II/6. A paraméterek és „odds”-ok intervallum becslése

Az X_j magyarázó változóhoz tartozó $\widehat{\beta_j}$ becsült paraméter a mintából származó pontbecslése a sokasági paraméternek. Ezért, szükséges a sokasági paramétert $1-\alpha$ megbízhatósági szinten tartalmazó konfidencia intervallum meghatározása. β_j paraméterés $1-\alpha$konfidencia szint mellett a keresett intervallum:

$$\widehat{\beta_j} \pm Z_{\left(1-\frac{\alpha}{2}\right)} * ASE\left(\widehat{\beta_j}\right)^6 \tag{2.15}$$

Hasonlóan az „odds”-ra gyakorolt multiplikatív hatás sokasági megfelelője az

$$e^{\left(\widehat{\beta_j} \pm \boldsymbol{Z}_{\left(1-\frac{\alpha}{2}\right)} * ASE\left(\widehat{\beta_j}\right)\right)} \tag{2.16}$$

intervallum része az adott megbízhatósági szinten.

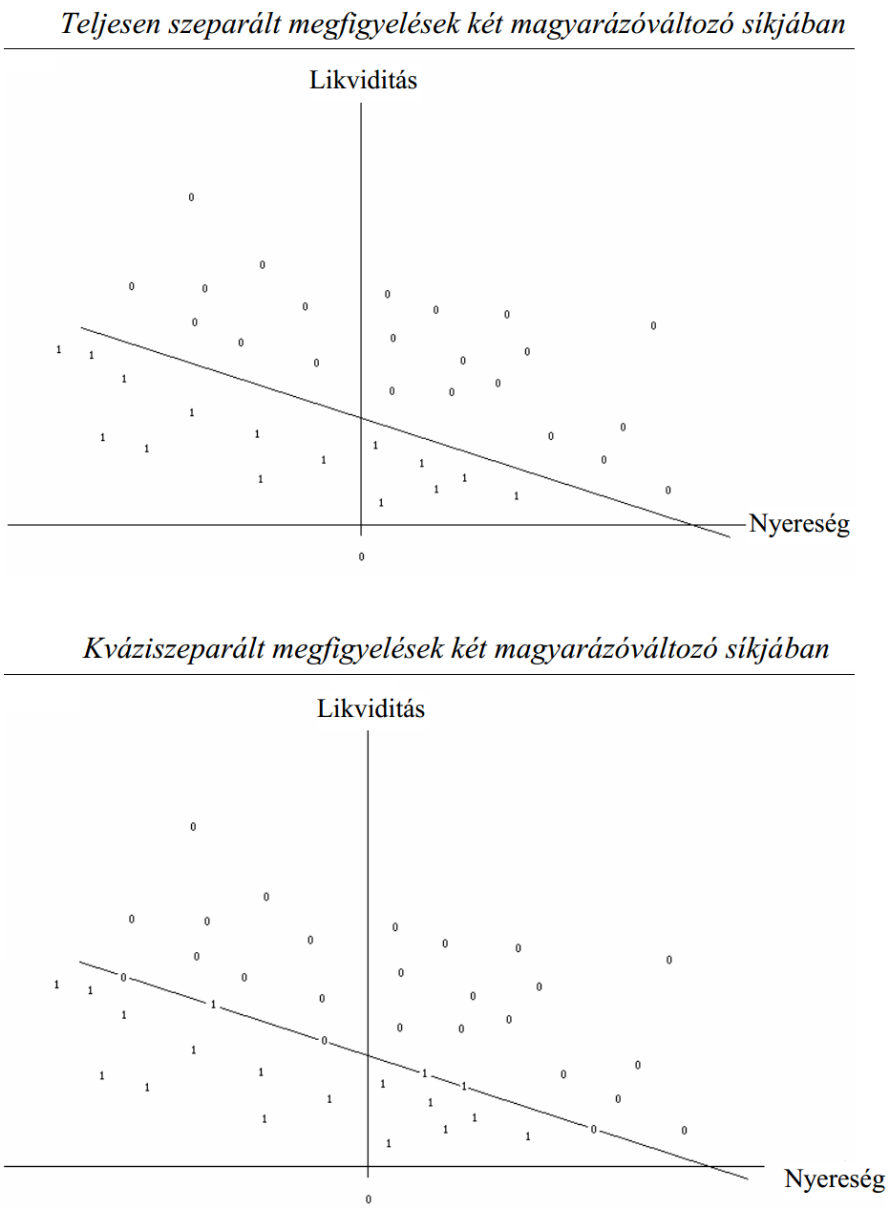
[6] ASE= Asimptotic Standard Error

II/7. Nem konvergáló ML- becslés problémája. Szeparáltság és kváziszeparáltság esete

Nem egyszer előforduló jelenség, hogy az iteratív algoritmus a paraméterek meghatározására nem eredményez véges értékű maximum likelihood becslést. Ilyenkor a programcsomagok általában a „final solution cannot be found" üzenettel jelzik, hogy az előre meghatározott iterációs lépésen belül[7]az algoritmus nem talált megoldást. Ilyenkor a modellezésbe bevont magyarázó változók valamelyike (és elég, ha csak egyetlen) szeparálja, vagy kváziszeparálja a mintát. Szeparáltságról beszélünk akkor, ha a magyarázó változó értékkészlete átfedésmentesen kettéosztja a mintát acélváltozó két kategóriája mentén. Erre jó például szolgálhat (Hajdu,2004) a vállalati nyereségesség példája, amelynél ha a mintában minden csődbement vállalat esetén negatív nyereség mutatható ki, illetve a fizetőképes vállatok esetében a nyereség pozitív értékű,akkor a nyereségesség értéke teljesen szeparálja a mintát. Ebben az esetben a nullaértékű nyereség teljesen szétválasztja az adatokat, hiszen egyik csoport esetén sem figyelhetjük meg. Ha a leírt példa csak annyiban módosul, hogy nulla nyereséget felvehetik mind a csődös, mind a fizetőképes cégek, akkor beszélünk kváziszeparáltságról. Két változó, a likviditás és a nyereség által kifeszített síkban (Hajdu, 2004) ábrája remekül szemlélteti a szeparáltság és a kváziszeparáltság esetét:

[7] Az SPSS alapbeállítása például 20 iterációs lépés

2. ábra Teljes szeparáltság és kváziszeparáltság két változó esetén (Forrás:Hajdu)



A kétváltozós példa arra is rávilágít, hogy előfordulhat, hogy a minta ugyan minden változó esetében külön-külön átfedéses (azaz sem teljes szeparáció, sem kváziszeparáció esete nem áll fent egy változóban), de a több változó által kifeszített térben található olyan sík (hipersík), mely az előbbi problémás esetek valamelyikét eredményezi.

A probléma jellemző előfordulása kisebb minták és kategóriaváltozók modellben történő alkalmazása esetén tapasztalható leginkább, amikor a kategória változók viszonylag sok kategóriával rendelkeznek. Ilyenkor, mivel a kategóriák kódolása dummy változókkal történik, könnyen előfordul, hogy az egyik dummy változó átfedésmentesen osztja fel a

megfigyeléseket a célváltozó kategóriái tekintetében. Ilyen esetben egyes kategóriák célszerű összevonása jelenthet megoldást, amiről még a későbbiekben részletesen szó lesz.

II/8. Változó szelekció

A modellezésbe bevont magyarázó változó paraméterének szignifikancia vizsgálata során derül ki, hogy mely változók relevánsak az előrejelzés szempontjából és melyek nem. Sajnos a nem releváns változókat nem lehet csak egyszerűen a modellből „elhagyni", mert a paraméterek ML becslése ezen változókkal együtt történt. Ezért a logisztikus regressziós modell kialakítása során fontos kérdés a releváns magyarázó változók körének kialakítása, azaz hogyan tudjuk a változóknak egy olyan részhalmazát kialakítani, melyek esetében mindegyik változó szignifikáns lesz. A probléma megoldásának széles körben elterjedt módja az ún. stepwise algoritmus használata. Az algoritmus úgy működik, hogy lépésenként mindig csak egy változót vonunk be a modellbe, vagy hagyunk el belőle. Döntésünk alapjául a likelihood arány teszt szolgál. Újabb változót csak akkor vonunk be a modellbe, ha hatására szignifikánsan (p_b) nő a likelihood függvény értéke, és csak akkor hagyunk el egy változót, ha annak hatására a likelihood függvény csökkenése nem szignifikáns (p_k)[8]. Aszerint, hogy a változószelekciót az összes változót tartalmazó modellből, illetve csak egy-egy változót tartalmazó modell irányából kiindulva kezdjük el, beszélünk *backward* illetve *forward* szelekcióról. Az algoritmus eredménye szempontjából kulcsfontosságú p_b és p_k értéke. A standard statisztikai programcsomagok alapbeállításként p_b=0,05 és p_k=0,1 értékeket adják meg. Lee és Koval (1997) vizsgálata megmutatta, hogy forward stepwise algoritmus használata esetén ezek az alapbeállítások túl szigorúnak bizonyulnak, használatukkal gyakran fontos magyarázó változót veszítünk el a modellből. Hosmer és Lemeshow (1999) hasonló célú kutatásában a p_b=0,15 és p_k=0,2 értékeket javasolja alkalmazni.

II/9. Illeszkedésvizsgálat

Az elkészült modell illeszkedésének vizsgálata a modellezési folyamat kritikus pillanata, itt dől el, hogy fáradozásaink, milyen mértékben voltak eredményesek. Ennek megfelelően részletesebben tárgyaljuk ezt a fejezetet. A logisztikus regressziós modell illeszkedésének

[8] (p_b)- beléptetési szignifikanciaszint, (p_k)- kiléptetési szignifikanciaszint

jóságát leíró statisztikák közül először azokat vesszük sorra, melyek speciálisan a logisztikus regresszió esetén alkalmasak az elkészült modell teljesítményének a mérésére.

Hosmer-Lemeshow-statisztika

A Hosmer-Lemeshow (HL) statisztika azt teszteli, hogy mennyire képes a modell az adott előrejelzési tartományokra megbecsülni a bekövetkezett események tényleges számát. A HL teszthez kapcsolódó táblázat értékei úgy keletkeznek, hogy az egyes esetek becsült bekövetkezési valószínűségeit növekvő sorba rendezzük, és az így keletkezett rangsort *k* (általában 10) egyenlő elemszámú csoportra bontjuk (kvantilisekbe (általában decilisekbe) rendezzük). Ezek után megvizsgáljuk, hogy az egyes csoportokba az egyes kategóriákból (1,0) hány megfigyelt (observed), és hány a regressziós becslés által várt (expected) eset tartozik. A Hosmer-Lemeshow statisztika értéke nem más, mint az erre a táblázatra alkalmazott Pearson féle *(k-2)*szabadságfokú χ^2statisztika. A HL statisztika kisebb értékei jelentik a jobb klasszifikációt (nagyobb besorolási pontosságot).
A HL statisztika hibája, hogy nagyon érzékeny a kialakított csoportok számára. Ha túl sok kategóriát alakítunk ki, akkor kevés lehet az egyes kategóriába eső ritka kategória eseteinek a száma, így a modell nehezen tudja pontosan megbecsülni, a HL statisztika pedig gyenge teljesítményt fog jelezni. Ha túl kevés a kategória, akkor a modell könnyen ad jó becslést, így a HL statisztika szerint mindig jó a modell előrejelző képessége.

Pszeudó- R^2 típusú mutatók

A lineáris regressziónál megszokott R^2 mintájára kerültek kifejlesztésre az úgynevezett pszeudó-R^2 teljesítmény mérőszámok, melyeknél a cél az volt, hogy ezen mutatók a [0,1] intervallumon vegyék fel lehetséges értékeiket és a nullához közeli értékek jelentsék a gyenge, míg az egyhez közeli értékek a jó illeszkedését a modellnek. Logisztikus regresszió esetében két, a programcsomagokban is gyakran előforduló, közkedvelt mérőszámot mutatunk be, melyek a csak konstans tagot tartalmazó nullmodell, illetve a mérendő *k* magyarázó változós modell likelihood-ját használják fel a kívánt cél eléréséhez. A **Cox-Snell** statisztika *n* elemű minta esetén az alábbi formula alapján számolható

$$R^2 = 1 - \left(\frac{-2LL_{null}}{-2LL_k}\right)^{\frac{2}{n}}. \tag{2.17}$$

A mutatóról elmondható, hogy jellemzően alacsonyabb értékkel jellemzi az illeszkedést, mint azt a hasonló teljesítményű lineáris regressziót leíró R^2 esetén várnánk. Ennek magyarázata az, hogy a Cox-Snell statisztika esetén a mutató értékének maximális értéke:

$$1 - LL_{null}{}^{\frac{2}{n}} \tag{2.18}$$

Vagy másképpen *p* megfigyelt esemény bekövetkezési arány esetén

$$1 - \left(p^p (1-p)^{(1-p)}\right)^2 \tag{2.19}$$

a maximum érték (Allison). Például P=0,1esetén a felső határértéke0,48 szemben a hagyományos R^2 esetén megszokott 1-gyel. A probléma megoldására kézenfekvő megoldás az, hogy a mutatót a felső határ arányában kifejezve alakítsuk át. Ennek a feltételnek tesz eleget az úgynevezett **Nagelkerke** statisztika

$$R^2 = \frac{1 - \left(\frac{-2LL_{null}}{-2LL_k}\right)^{\frac{2}{n}}}{1 - (-2LL_{null})^{\frac{2}{n}}} \tag{2.20}$$

A Naglerkerke statisztika számolásából következik, hogy értéke nagyobb a Cox-Snell statisztika értékénél adott modell esetén.

Brier-score, logaritmikus- score

Az úgynevezett Brier-score bemutatásával áttérünk az általános célú, nemcsak a logisztikus regressziós modell esetén alkalmazható, teljesítménymérésre alkalmas mutatószámok területére. A mutató számolása a következő formula alapján történik

$$BS = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(1|x_i)\right)^2 \tag{2.21}$$

Itt y_i = 1, ha bekövetkezik az esemény, és y_i = 0 ha nem, $\hat{f}(1|x_i)$ pedig annak a becsült valószínűsége, hogy az i objektum a bekövetkezők közé tartozik. A mutató értékének elméleti minimuma 0 (tökéletes modell esetén, ami a nem bekövetkezők esetén 0 bekövetkezési valószínűséget, a bekövetkezők esetén 1-et becsül), maximuma 1 (épp ellentétes besorolás esetén). Ezen mutatóval egyező elven működik a logaritmikus score:

$$LS = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \ln \hat{f}(1|x_i) + (1-y_i)\ln\left(1-\hat{f}(1|x_i)\right)\right) = -\frac{1}{N}\sum_{i=1}^{N}\ln\left(|y_i + \hat{f}(1|x_i) - 1\right) \quad (2.22)$$

A mutató 0 közeli értékei jelzik a modell jó teljesítményét. A Brier és a logaritmikus score is az előrejelzési hibát mérik, ezért a 0 közeli értékek jelentik a nagyobb besorolási pontosságot.

Konfúziós mátrix

Klasszifikációs célú prediktív modellek széleskörűen elterjedt teljesítménymérési eszköze a konfúziós mátrix, vagy más néven klasszifikációs tábla. A könnyebb megértés érdekében tegyük fel, hogy elkészült modellünket hitelezési döntéshez kívánjuk felhasználni, ahol azt kívánjuk előrejelezni, hogy egy adott vállalat fizetőképesség szempontjából kialakított „túlélő" és „csődös" kategória melyikébe fog esni. A döntés egy adott vállalat esetében úgy történik, hogy a vállalat modell által szolgáltatott függvényértékét (nulla és egy közötti szám) viszonyítjuk egy előre definiált döntési küszöb értékhez (cut-off point). Amennyiben ez az érték nagyobb egyenlő, mint a cut-off, akkor csődősnek tekintjük, egyéb esetekben pedig túlélőnek. Az adott mintán és cut-off mellett az összes értékelendő vállalaton elvégezve a leírt besorolást, majd összevetve a valóságos kategóriába való tartozással kapjuk a konfúziós mátrixot:

1. Táblázat. Konfúziós mátrix

		Tényleges csoport	
		csődös	túlélő
Előrejelzett csoport	csődös	TP helyes besorolás (csőd)	FP elsőfajú hiba
	túlélő	FN másodfajú hiba	TN helyes besorolás (túlélő)

Ahol TP= True positive, TN= True negative, FP= False positive, FN= False negative.

A táblázatból látható, hogy a helyes besorolások száma= TP+TN, míg a téves besorolások száma=FP+FN. Elsőfajú hibának azt tekintjük, amikor egy ténylegesen túlélő vállalatot tévesen csődösnek minősítünk (FP), míg másodfajú hiba esetén egy csődös vállalatot minősítünk tévesen túlélőnek (amennyiben H_0: A vállalat túlélő). A mátrix elemeiből számos mutatószám képezhető, melyekből fontosságuk miatt kettőt emelnénk ki. Az első az úgynevezett TPR (true positive rate), melyet találati érzékenységnek (sensitivity) is neveznek

$$TPR = \frac{TP}{TP + FN} \tag{2.23}$$

A kifejezés megmutatja, hogy adott cut-off mellett modellünk hány százalékát képes helyesen felismerni (besorolni) a csődős vállalkozásoknak. A másik fontos mutatószám az FPR (false positive rate), melyet a következőképpen írhatunk fel

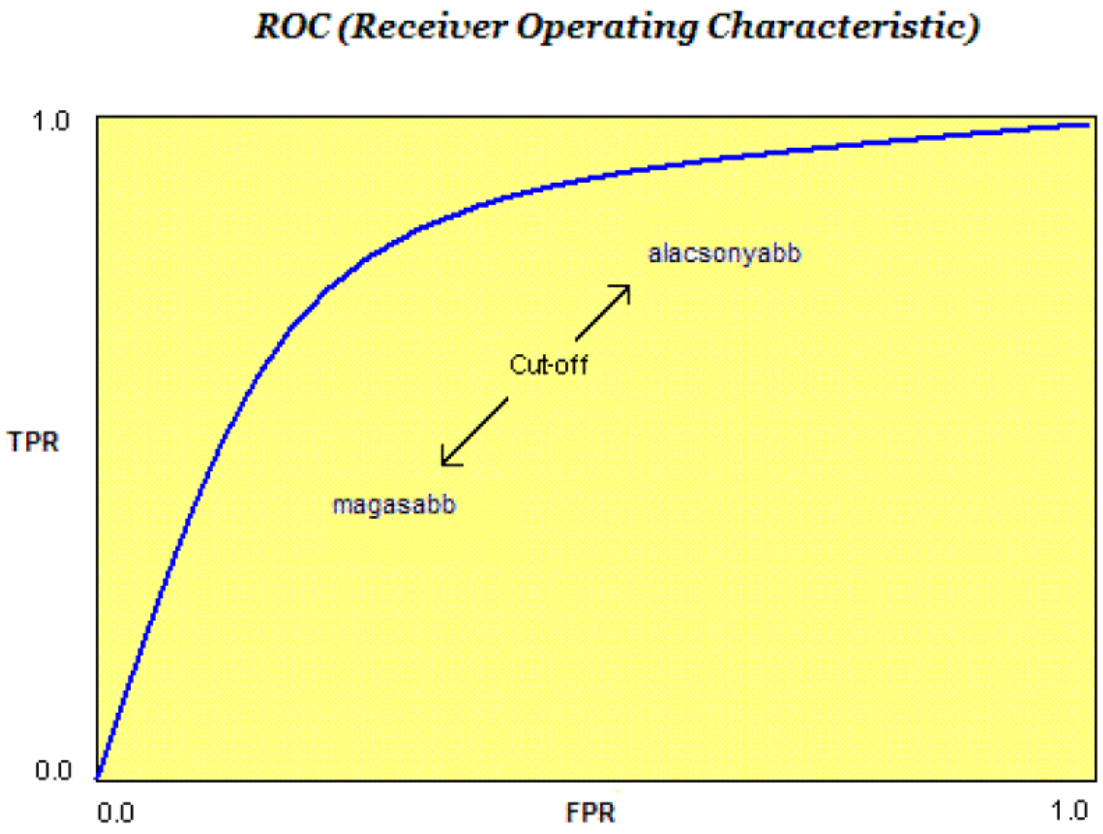
$$FPR = \frac{FP}{FP + TN} \tag{2.24}$$

és megmutatja, hogy adott cut-off mellett a túlélő vállalatok hány százaléka lesz tévesen csődösnek minősítve.

ROC- görbe

A ROC (Receiver Operating Characteristic) görbét először a II. világháborúban használták radarjelzések elemzésére. A módszer ötlete abból a felismerésből ered, hogy a konfúziós mátrix elemei egy adott cut-off mellett értelmezhetők, a cut-off érték megváltoztatásával a mátrix elemei is új értékeket vesznek fel. Úgyis mondhatnánk, hogy ahány cut-off értéket veszünk, annyi konfúziós mátrixot kapunk. A ROC görbe egy ábrába sűrítve mutatja meg számunkra, hogyan változik a TPR és FPR mutatók értéke, amint a cut-off értéket nulla és egy között mozgatjuk:

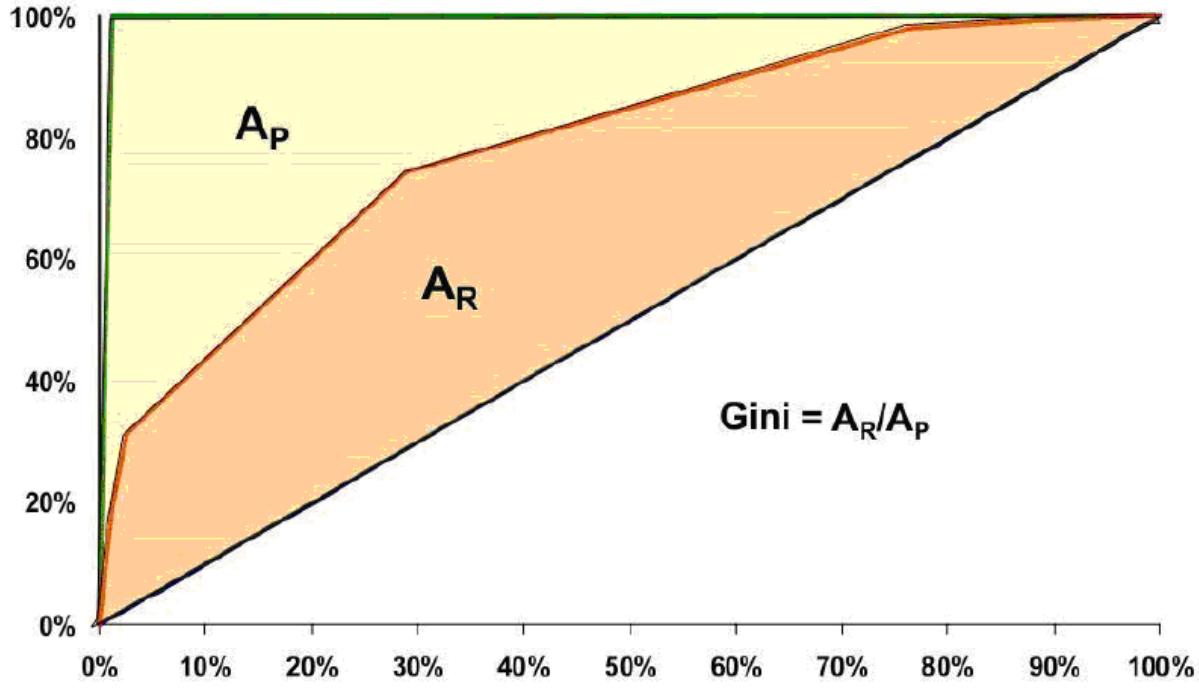
3. ábra. ROC görbe



Az ábrát tanulmányozva könnyű belátni, hogy a tökéletesen klasszifikáló (előrejelző) modell esetében a kék ROC görbe az ábra téglalapjának bal felső sarkába simulva végigköveti a baloldali függőleges és a felső vízszintes szakaszt, míg a klasszifikáló erővel egyáltalán nem rendelkező modell görbéje a téglalap (0,0) pontból kiinduló átlójára illeszkedik. A valóságos modellek jellemző ROC görbék természetesen valahol a két

szélsőséges helyzet között szoktak elhelyezkedni. A ROC görbe segítségével az elkészült modellről egy globális képet kapunk, hiszen bármely cut-off érték esetén az ábráról könnyen leolvasható és számolható a konfúziós mátrix valamennyi eleme. Felmerül a kérdés, hogy a lineáris regressziónál megszokott R négyzet mintájára alkotható-e egy olyan mutatószám a ROC görbe segítségével, melynél a tökéletesen illeszkedő modell esetén a mutató értéke egy, illetve az előrejelző erővel nem bíró modell esetén ez az érték nulla. A leginkább elterjedt megoldás az úgynevezett Gini koefficiens használata, melynél a görbe és az átló közötti területet (A_R) hasonlítjuk a bennfoglaló derékszögű háromszög területéhez (A_P) az alábbi ábrán látható módon.

4. ábra. Gini koefficiens



A területhányados alapján könnyen belátható, hogy a Gini mutató értéke 1 a perfekt modell, és 0 a véletlenszerűen előrejelző modell esetén[9].

[9] Tökéletes modell esetén $A_R=A_P$, a véletlenszerűen előrejelző modell esetén $A_R=0$

Konklúzió

A bemutatásra kerülő illeszkedési mérőszámok közül a nem logit-specifikus Brier- és logaritmikus score, valamint a Gini-koefficiensa széles körben elterjedtek. Ennek az oka, hogy ezek a mutatók minden további nélkül számolhatók a fejlesztési (training) mintától különböző tesztmintára is. Ezzel szemben a többi, logit-specifikus mérőszámot a programcsomagok alapértelmezésben csak a fejlesztő mintára számolják ki. Grafikus tulajdonságai miatt könnyen interpretálható ROC-görbe a belőle számolható Gini-koefficienssel a gyakorlati modellezés legkedveltebb mérőszáma.

III.A modellezési adatbázis kialakításának szempontjai

III/1Reprezentativitás, EPV

A modellezés alapjául szolgáló adatállomány összeállításakor a legfontosabb szempont az adatbázis reprezentativitása. Az adatállománynak reprezentálni szükséges azt a populációt, melyre a kialakításra kerülő modellalkalmazásra kerül.

Nagyon nagyméretű adatállományok (több millió eset több száz változóval) esetében, még a mai számítási teljesítmények mellett is előfordulhat túl hosszú futásidő. Ebben az esetben a megoldást a mintavétel jelentheti, ami a leggyakrabban az egyszerű véletlen minta leválasztásával érhető el. Ha az adatállomány mérete várhatóan ilyen problémát nem okoz, akkor a teljes adatbázis használható kiindulásként.

A többváltozós statisztikai modellek kialakítása esetében problémát okozhat, ha túl kevés a megfigyelések száma a lehetséges magyarázó változók számához képest (EPV= Events Per Variable). Logisztikus regresszió esetében az alacsony EPV értékek mellett a paraméterbecslések torzítottá válnak és megnő a szélsőséges maximum likelihood becslés esélye is (Peduzzi, Concato, Kempner,1996). Peduzzi, Concato, Kempner különböző EPV értékek mellett elvégzett Monte Carlo szimulációja alapján, az EPV=10 értéket javasolja minimumkritériumként használni a modellezési adatbázis kialakítása során.

III/2 Modellezési adatbázis kis és kiegyensúlyozatlan minták esetén

A logisztikus regresszió paramétereinek becslésére a korábbiakban már tárgyalt maximum likelihood módszer segítségével történik. Az aszimptotikusan kedvező tulajdonságokkal[10] rendelkező maximum likelihood módszer azonban hagyományosan nagymintás módszernek tekinthető. Ennek következtében kisminták esetében a logisztikus regresszió paraméterbecslése torzítottá válik. A torzítás nagysága azonban a mintaméret növelésével csökken, n=200 felett már nem számottevő (Schaufer, 1983).

További probléma az ún.kiegyensúlyozatlan minták esete. A gyakorlatban sokszor fordul elő, hogy a logisztikus regresszió függő (cél) változójának egyik kategóriája jóval kisebb

[10] Konzisztens, aszimptotikusan hatásos, határeloszlása normális (Hunyadi és Vita, 2002)

elemszámú (< 5 %), mint a másik. Ebben az esetben beszélünk az ún. kiegyensúlyozatlan mintáról és az alacsony elemszámú kategóriát hívjuk ritka eseménynek. A ritka esemény következtében a paraméterekre torzított és magas varianciájú ML becslést kapunk tekintet nélkül a mintanagyságra (Hajdu, 2004). A torzítás egyirányú, a regresszió segítségével meghatározott valószínűségi pontbecslés egy adott egyed esetében alulbecsüli a sokasági valószínűséget. A probléma kezelésére általánosan elterjedt megoldás további ritka események csatolása az adatbázishoz, melynek további kedvező hatása a becslés standard hibájának csökkenése is. Amennyiben nincs lehetőség további ritka egyedek csatolására, akkor célszerű a „nem ritka" kategória egyedei közül, megfelelő mintavételi stratégia mentén, elhagyni egyeseket[11] (King-Zeng, 2001). Ez utóbbi esetben törekedni kell az 50-50% kategóriaarány elérésére, mely optimális a paraméterek standard hibája szempontjából (Hajdu, 2004). Akár a további egyedek csatolása, akár az egyedek elhagyása mellett döntünk, az adatmanipuláció után az eredeti ritka esemény arány meg fog változni. Amennyiben ismerjük a sokasági ritka esemény arányt (*prior*), és a mintabeli arány ettől eltér, akkor szükségessé válik ún.torzítás csökkentő módszerek használata a megfelelő valószínűségbecslés érdekében, melyeket a későbbiekben részletesen tárgyalni fogunk.
Mind a kis minta, mind a kiegyensúlyozatlanság kérdésére lehetséges választ jelent egy nem aszimptotikus módszer az ún. egzakt logisztikus regresszió (ELR) használata. Az ELR eljárás a regressziós paraméterek elégséges statisztikáinak az egzakt, feltételes, permutációs eloszlásán alapuló módszer (Hajdu, 2004). A módszer ismertetése a disszertáció keretein túlnyúlik, csak annyit említünk meg, hogy széleskörű elterjedését a gyakorlatban, különösen már közepes méretű adatbázisokon is tapasztalható túlzott számításigényessége akadályozza.

III/3 Adatbázis particionálása

Amennyiben előrejelző modellünket a teljes mintán építjük fel, és a besorolási pontosságát is ugyanezen a mintán ellenőrizzük, számíthatunk arra, hogy modellünk magyarázó erejét (illeszkedését) kedvezőbbnek fogjuk megítélni, mintha az ellenőrzést egy másik, az előzőtől független mintán hajtottuk volna végre, azaz alulbecsüljük a téves besorolás valószínűségét. Kellően nagy minta esetén a mintát, ahogyan erre az illeszkedésvizsgálati

[11] A módszert szokásosan *case-control* módszerként említi a szakirodalom

résznél már utaltunk,két részre particionálják, és az egyik minta szolgál a modellépítés céljaira (tréning-training), míg a másik a modell prediktív erejének ellenőrzésére (teszt-test). A leggyakrabban alkalmazott felosztási arányok a 80-20% és a 70-30% a tréning és teszt minta tekintetében. Fontos megjegyezni, hogy a mintafelosztás arányára szintén nincs egyértelmű iránymutatás a szakirodalomban, ami azért fontos kérdés, mert Hu(1999) eredményei szerint a felosztás arányának változtatása akár 2-3 százalékpontos változást is eredményezhet a modellek validációs mintán mért besorolási pontosságában.A módszer hátránya, hogy a kisebb minta következtében mintainformációt veszítünk, a paraméterbecslések standard hibája növekszik. Kisebb minták esetén ezért előfordulhat, hogy egy adott magyarázó változó meghatározott szignifikancia szinten már nem szignifikáns, rontva ezzel a modell teljesítményét a teljes (felosztás nélküli) adatbázison épített modellhez képest. Az esetleges romlást felfoghatjuk úgy is, mint annak a tudásnak az „árát", melyet a modell teljesítményének objektívebb ismerete érdekében vállalunk fel. Egy adott modell futtatása esetén a felosztás tehát mérlegelés kérdése. Amennyiben viszont egy adott előrejelzési problémára több egymással versenyző modelltípust is használunk a felosztás szükségessé válik az alternatív modellek teljesítményének megfelelő összehasonlíthatósága érdekében.

Másik elterjedt módszer az úgynevezett „Jackknife", vagy más néven „leave one out" technika, amely a paraméterek becslésénél mindig egy esetet elhagy a mintából és a megmaradt adatokon történik a paraméterek becslése. Az így elkészült modell kerül ellenőrzésre a kihagyott eseten, azt vizsgálva, hogy vajon jól sorolja-e be azt. Az eljárás így folytatódik mindaddig, míg az összes eset sorra nem kerül. A végleges paraméterbecslés az egyedi regressziók becsléseinek számtani átlagaként áll elő. A módszer hátránya, hogy számításigényes, hiszen egy n elemű mintán n-szer kell a paramétereket becsülni.

IV.Hiányzó értékek és kezelésük

A standard statisztikai módszerek, így a prediktív modellek is, adathiányt nem tartalmazó, adatbázisok elemzésére lettek kifejlesztve. Modellépítés során azonban gyakran találkozunk hiányos adatmátrixokkal. Miért okoz ez problémát? Azért, mert a legtöbb statisztikai szoftver a hiányzó adatokat alapbeállításként úgy kezeli, hogy törli az elemzésből azt a megfigyelést, amely akár egyetlen változóban is adathiányos (listwise deletion). Így gyakran nagy hányada az adatmintának kimarad a modellezésből, csökkenő esetszámot és ez által modelljeink romló statisztikai tulajdonságait eredményezve. Legtöbb esetben sérül a minta véletlenszerűsége is, melynek következtében paraméterbecslések torzítottá válnak. Az alábbiakban röviden áttekintjük a jellemző adathiány típusokat és a kezelésükre kifejlesztett módszereket.

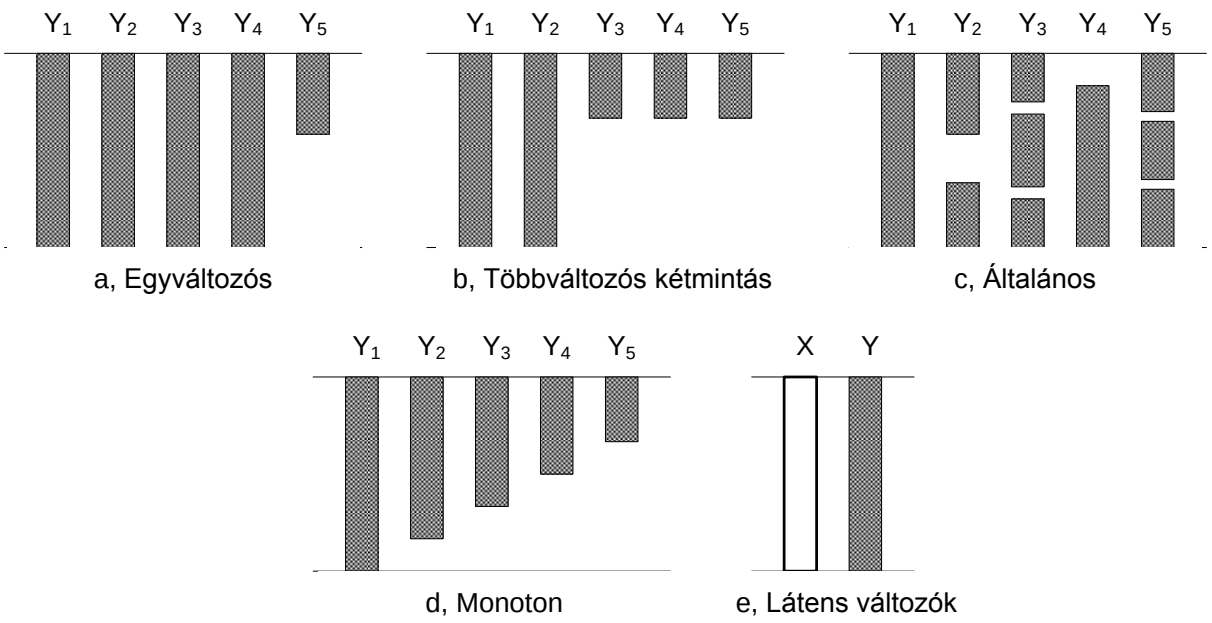
Az osztályozás egyik szempontja az *adathiány mintázata*. A mintázat az írja le, hogy a hiányzó adatok milyen módon hiányoznak az adatmátrixban. A másik osztályozási szempont az *adathiány mechanizmus*, amely a hiányzás és az adatbázisban szereplőváltozók értékei közötti kapcsolatot vizsgálja.

IV/1 Hiányzóadatok mintázata

Legyen $Y = (y_{ij})$ egy (n x K) általános adatmátrix, hiányzó adatok nélkül, amelynek i -dik sora $y_i = (y_{i1}, \ldots, y_{iK})$, ahol y_{ij} az Y_j változó értéke az i –dik egységnél. Hiányzó adatok esetén legyen $M = (m_{ij})$ az adathiány indikátor mátrix (Oravecz,2008), ahol $m_{ij} = 1$, ha y_{ij} hiányzik és $m_{ij} = 0$, ha y_{ij} megfigyelt. Az M mátrix definiálja az adathiány mintázatot. Az alábbi ábrán a nem hiányzó értékeket sötéttel jelölve láthatunk néhány példát az adathiány mintázatokra.

5. ábra.Adathiány mintázatok (Forrás:Oravecz)

.



Egyváltozós adathiány

Az 5a) ábrán látható az az esetet, amikor csak egyetlen változóban vannak hiányzó értékek, a többi változó hiánytalan. Ilyen mintázata lehet például egy kérdőíves adatfelvétel eredményének, ahol az adathiányos változó a megkérdezettek jövedelmét tartalmazza, és a válaszadók egy része erre a kérdésre nem volt hajlandó válaszolni, a többi kérdésre viszont igen. Másik példa lehet egy olyan adatbázis, ahol az egyetlen hiányzó értékeket tartalmazó változó azt az információt kódolja, hogy egy adott ember hány éves volt, amikor megkezdte sorkatonai szolgálatát. Értelemszerűen a nők és a sorkatonai szolgálatot sosem teljesítők esetében adathiány fog mutatkozni ebben a változóban.

Többváltozós két mintás

Egy másik általános minta, amikor az előző példában szereplő egyetlen adathiányos változó (Y_K) helyett több adathiányos változónk van $(Y_{j+1},\ldots,Y_K)$, ahol mindegyik egyformán megfigyelt, vagy hiányzik ugyanazokra az esetekre. (Lásd a 4b ábrát.)

Ezt a mintázatot eredményezheti például a kérdőíves felméréseknél az egység szintű nemválaszolás[12] jelensége. Az egység szintű nemválaszolás előfordulhat azért, mert a kiküldött kérdőívet meg sem kapta a címzett, vagy megkapta, de megtagadta a válaszadást. Ekkor a kérdőívben szereplő változók lesznek az adathiányos változók. A teljes, adathiányt nem tartalmazó változók a minta tervezéséhez használt változók lesznek, amelyek mind a válaszadók, mind a nemválaszolók esetében előzetesen ismertek egy listáról (Oravecz,2008). Másik példa lehet egy olyan adatbázis, ahol vállalatok pénzügyi beszámolóiból képzett mutatók szerepelnek. A vállalatok, méretüknek megfelelően, készítenek éves és egyszerűsített éves beszámolót. Az egyszerűsített beszámoló számos információt nem tartalmaz, amit az éves beszámoló viszont igen. Így például a fizetett kamatok/pénzügyi ráfordítások mutató csak az éves beszámolót készítő vállalkozások esetében lesz kitöltött, mert a fizetett kamatokat az egyszerűsített beszámoló nem tartalmazza.

Általános mintázat

Általános mintázatról beszélünk, ha az adathiány mintázata látszólag semmiféle szabályszerűséggel nem rendelkezik. (Lásd 4c ábra.).

Monoton adathiány

Monoton adathiányra keletkezésére jó példa a longitudinális adatfelvételek során bekövetkező lemorzsolódások esete, azaz amikor egy vagy több megfigyelési egység kiesik a mintából[13] Ilyen esetről beszélünk, ha a háztartás panel esetén a család külföldre költözik, vagy klinikai kísérleteknél más gyógyszerek hatása, vagy egyéb betegség miatt a beteg nem tud tovább részt venni a kísérletekben (Oravecz, 2008). De a szolgáltató nagyvállalatok esetében tapasztalható ügyfél lemorzsolódást (churn) is ugyanide sorolhatjuk. Monoton adathiány esetén a változókat el lehet úgy rendezni az adatmátrixban, hogy minden $Y_{j+1},\ldots,Y_K$ hiányos, ha Y_j hiányzos (Lásd 4d ábra.), azaz a változónkénti adathiány mértékének a változása monoton legyen változóról változóra haladva.

[12] Amennyiben az adatbázisból egy-egy megfigyelés teljesen hiányzik teljes (vagy egység szintű) nemválaszolásról (unit nonresponse) beszélünk.

[13]A longitudinális adatfelvételek különböző időszakokban gyűjtenek be adatokat ugyanazon megfigyelési egységekről.

Látens változók

A nem megfigyelhető látens változókat is felfoghatjuk adathiány problémaként, csak ezeknél a látens változóknál speciálisan minden megfigyelési érték hiányzik. Az 4e) ábrán az X jelenti a látens változók csoportját, ahol minden érték hiányzik és Y pedig a teljesen megfigyelt változók csoportját. Ekkor természetesen bármiféle elemzéshez különböző feltételezésekkel kell élnünk.

IV/2 Adathiány keletkezésének típusai

Az adathiány kezelésére választandó módszer kiválasztása függ attól, hogy milyen módon keletkeztek az adathiányok, azaz mi az a mechanizmus, amelynek következtében kialakultak.

Little és Rubin (1987) az adathiány három alapvető esetét különbözteti meg, attól függően, hogy milyen a kapcsolat a hiányzás és az adatbázisban lévő változók értékei között.

Ennek leírására legyen továbbra is az $Y = (y_{ij})$ a teljes adatmátrix és az $M = (m_{ij})$ az adathiány indikátor mátrix. Az adathiány mechanizmus jellemezhető az M adott Y melletti feltételes eloszlásával, az $f(M|Y,\Phi)$-vel, ahol Φ ismeretlen paramétereket jelöl.

Teljesen véletlenszerű adathiány

A teljesen véletlenszerű adathiány (Missing Completely at Random (MCAR)) esetén a teljes adatállománnyal rendelkező részek és a hiányzó adatokat tartalmazó megfigyelések ugyanabból az eloszlásból származnak. A hiányzás tehát nem függ az Y értékétől, sem a megfigyelt (komplett), sem a hiányzó adatokkal rendelkező változók értékétől, azaz:

$$f(M|Y,\Phi) = f(M|\Phi), \text{ minden } Y,\Phi \text{ esetén.} \qquad (4.1)$$

Egy változóban tapasztalható adathiány esetén ez annyit jelent, hogy az adathiány nem függ az adatbázis komplett változóitól, sem pedig önmagától az adathiányos változótól. Felvetődhet a kérdés, hogy a MCAR megléte tesztelhető-e? Annak eldöntése, hogy a hiányzások függenek-e a többi változótól nem nehéz. Elterjedt módszer logisztikus regresszió alkalmazása a komplett változókkal, mint prediktorokkal és hiányos változó

indikátor változójával, mint célváltozóval. Szignifikáns koefficiensek megléte a regresszióban jelzi a kapcsolatot (függést) a komplett változókkal és ezzel a MCAR feltételezés fenntarthatatlanságát. Annak eldöntésére viszont, hogy a hiányos változó önmagával magyarázható-e, sajnos nincs mód. Ennek eldöntéséhez ismernünk kellene magukat a hiányzó értékeket. Ennek következtében a legtöbb elemzésre kerülő adatbázis esetében a MCAR feltételezés nem elégíthető ki bizonyosan. Ez alól kivételt jelenthet a tervezett hiányzások (*missing by design*) esete (Graham,1996). Példaként tételezzük fel, hogy egy klinikai kutatásban fontos változók származnak a páciensek CT vizsgálatából. A vizsgálat költséges mivolta miatt a kutatók úgy döntenek, hogy a pácienseknek csak egy véletlenszerűen kiválasztott 60%- nál végzik el a vizsgálatot. A maradék 40% esetében az így keletkező adathiány már kielégíti a MCAR feltételezést. A véletlenszerűség következtében a MACR adathiánnyal jellemezhető adatbázisok komplett esetein (listwise deletion) végzett regressziók a paraméterek torzítatlan becslését eredményezik. A standard hibák pedig az elemzésből kimaradt esetek számával arányosan növekszenek.

Véletlenszerű adathiány

A MCAR feltételezésnél gyengébb feltételezés a véletlenszerű adathiány (Missing at Random (MAR)) esete. Ebben az esetben a hiányzó adatokat tartalmazó egységek eltérnek a hiánytalan adatokkal bíró egységektől, de a hiány előre jelezhető az adatbázis más változói segítségével. Az adathiány tehát más változókkal kapcsolatban van, de azzal a változóval, amelyikben a hiányzás felmerül, nincs közvetlen kapcsolatban.
Legyen $Y_{megfigyelt}$az a része *Y*-nak amelyben nincs adathiány és $Y_{hiányzó}$ az a rész, amelyben van adathiány. A véletlenszerű adathiány tehát az jelenti, hogy:

$$f(M|Y,\Phi) = f\left(M\middle|Y_{megfigyelt}\ \Phi\right), \text{ minden } Y_{hiányzó}, \Phi \text{esetén.} \qquad (4.2)$$

Az, hogy a hiányzó értékeket tartalmazó változók esetében a hiányzó értékek függetlenek-e önmaguktól a hiányzó értékeket tartalmazó változóktól, hasonlóan a MCAR esethez, itt sem tesztelhető. Egyváltozós adathiánynál például a magas jövedelműek esetében nagyobb valószínűséggel fog hiányozni a jövedelem értéke, mint az alacsonyabb jövedelműeknél. Ezt a jelenséget azonban magában az adatmátrixban semmilyen módon nem tudjuk detektálni. A MAR feltételezés megtartására szerencsére ebben az esetben is kínálkozik

mód (Allison,2013). Amennyiben a modellezési adatbázisba bevonunk a hiányos változóval erősen korreláló változókat, úgy a hiányos változó reziduális függése önmagától jelentősen csökkenthető, vagy eliminálható a prediktorok viszonylatában. Ez a konkrét jövedelmi példában a kor, iskolai végzettség, beosztás, nem változók bevonását jelentheti például.

Nem véletlenszerű adathiány

Nem véletlenszerű adathiány (Not Missing at Random (NMAR)) esetén az adathiány nem véletlenszerű, és önmagában más változókkal sem becsülhető, mert közvetlenül az adathiányt tartalmazó változóval is kapcsolatban van. Az *M* eloszlása tehát függ az *Y* hiányzó értékeitől is. Ez az adathiány legproblematikusabb formája (Oravecz, 2008). Ebben az esetben az adathiányt generáló mechanizmust a teljes modellezési folyamat részeként, abba beágyazottan külön szükséges modellezni. Itt a nehézség forrása az, hogy az NMAR mechanizmusa problémáról problémára változik, nem tipizálható. További gondot jelent, hogy az adatokban nincs információ arra vonatkozóan, hogy milyen modell választása lenne a legcélszerűbb, továbbá nem léteznek statisztikák, amelyek a modell illeszkedését írnák le. Ráadásul az eredmények nagyban függnek a választott modell típusától (Little and Rubin, 2002). A probléma kezelésére kezdetben ígéretes megoldásnak tűnt Heckman kétlépcsős probit modellje (Heckman,1979), amelyet később jelentős kritikák értek a módszer érzékeny és megbízhatatlan mivolta miatt (Chen, Asterbo, 2001). Manapság egyre szélesebb körben kezd kialakulni az a nézet, hogy a NMAR kezelésére a leghatékonyabb megoldás a pótlólagos adatgyűjtés jelentheti (Hand, Henley, 1993). Ha erre nincs mód, megfontolandó a modellezési feladat átfogalmazása és szűkítése a nemhiányzó adatokra. A jövedelemi példában ez egy olyan modell elkészítését jelentheti, ami kihagyja a magas jövedelemmel rendelkezők megfigyeléseit az adatmintából.

A vázolt nehézségek tükrében nem meglepő, hogy az elérhető statisztikai szoftverek esetében a hiányzó értékek kezelésére kifejlesztett eljárások az MCAR, illetve a MAR feltételezéseken nyugszanak. Továbbiakban az adathiány kezelésének módszereit, ezen belül is elsőként a hagyományos módszereket tekintjük át.

IV/3. Hiányzó adatok kezelésének hagyományos módszerei

Változó elhagyása

Ha egy független változóban az adathiányos egyedek hányada számottevő (>50%) megfontolandó a változó elhagyása a modellezésből. Ez leginkább akkor nem jelent problémát, ha léteznek az adatmátrixban az adathiányos változóval erősen összefüggő/korreláló hiánytalan változók is, általuk a törölt változó információtartalmának jelentős része reprezentálódik.

Adathiányt tartalmazó esetek elhagyása

Ennek a már a fejezet elején említett módszernek (Listwise vagy casewise deletion) nagy előnye az egyszerűsége. Csak ismétlésképpen, ilyenkor az eset, mely akárcsak egy változóban is adathiányt mutat törlésre kerül az adatmátrixból. MCAR adathiány esetén a paraméterek torzítatlanul becsülhetők, hiszen a hiányzó adatokat tartalmazó esetek az összes eseten belüli véletlenszerű almintának tekinthetők. A módszer további kedvező tulajdonsága, hogy amennyiben a prediktorok adathiánya független a célváltozótól a paraméterek továbbra is közelítőleg torzítatlanul becsülhetők MCAR és MAR feltételek sérülése esetén is különböző regressziós modellek[14] alkalmazásakor (Little, 1992).
A módszer hátránya a törlések következményeként csökkenő elemszám, mely különösen általános adathiány mintázat és nagyszámú változó esetén akár extrém mértékű is lehet[15]. A csökkenő elemszám kedvezőtlen hatása továbbá a standard hibák növekedése is.

Páronkénti törlés-rendelkezésre álló esetek

Lineáris modellek (lineáris regresszió, faktor analízis, SEM modellek, stb) esetén nyújt népszerű alternatívát a rendelkezésre álló esetek elemzésének módszere. A lineáris modellek paraméter becsléseinél központi jelentőségű a páronkénti kovariancia-variancia struktúra és az azokkal történő műveletek. A kétváltozós kovariancia vagy korreláció

[14] Lineáris, logit, Poisson, Cox, stb.
[15] Ha elképzeljük azt a hipotetikus helyzetet, hogy száz darab változónk van egyenként egy százalékos adathiánnyal, úgy hogy egy adott esetnél csak egy ,azaz egy változóban található adathiány akkor listwise módszer az egész adatbázist törli, egyetlen egy esetet sem hagyva elemzési célra.

számolásakor a *Páronkénti törlés* (*Pairwise deletion*) technika használatával a vizsgált két változó esetén felhasználásra kerül minden nem adathiányos megfigyeléspár. Ennek előnye, hogy több információ kerül felhasználásra, mint a *listwise deletion* esetén. A módszer MCAR esetén a paraméterek torzítatlan becslését eredményezi, azonban MAR esetén a becslések torzítottá válnak. A technika további jelentős hátránya, hogy előfordulhatnak olyan adatszerkezetek, melyeknél a *pairwise deletion* alkalmazása esetén a korrelációs mátrix nem lesz pozitív definit[16], lehetetlenné téve lineáris modell becslését. A következő egyszerű példa (Oravecz,2008) segítségével könnyen megérthetjük a problémát. Az alábbi táblázat 3 változóban összesen 12 megfigyelést tartalmaz és a „?" hiányzó adatot jelent:

2. Táblázat. Páronkénti törlés (Pairwise deletion)

Y_1	1	2	3	4	1	2	3	4	?	?	?	?
Y_2	1	2	3	4	?	?	?	?	1	2	3	4
Y_3	?	?	?	?	1	2	3	4	4	3	2	1

Ebben az adattáblában az elérhető adatpárokat használva $r_{12} = 1$, $r_{13} = 1$, $r_{23} = -1$. Ezek a becslések nem jók, mert $\rho_{12} = \rho_{13} = 1$ –ből az következik, hogy $\rho_{23} = 1$, nem -1.

IV/4. Imputációs módszerek

Az imputáció alapú eljárások közös jellemzője, hogy valamilyen módon a hiányzó értékek helyére utólag adatot helyettesítenek, így az adatbázis a standard statisztikai eszköztárral elemezhetővé válik. A továbbiakban az egyszerűtől az összetettebbek felé haladva sorra vesszük a legelterjedtebb technikákat.

Középértékkel történő pótlás

Az adathiányos folytonos változóban az adathiányos helyekre a változó megfigyelt értékeinek az átlagával, mediánjával, vagy móduszával hajthatjuk végre az adatpótlást. Ezek közül a szoftverek körében is a legelterjedtebb az átlaggal történő imputáció. A

[16] A paraméterek becsléséhez szükséges inverz mátrix ebben az esetben nem számolható.

módszer egyszerűsége mellett számos árnyoldallal is rendelkezik. Az átlaggal imputált adatokon a paraméterek csak torzítottan becsülhetők (Haitovsky, 1968), és emellett az elemek változékonysága is alulbecsült. Ez utóbbi javítható, ha a megfigyeléseket homogénebb csoportokra bontjuk és csoportokon belüli részátlagokkal imputálunk, de a standard hibákat és a becslések konfidencia intervallumát még így is alulbecsüljük.

Lineáris regresszióval történő pótlás

A középértékkel történő imputálásnál lényegesen jobb eredményt érhetünk el a lineáris regresszióval történő pótlással. Ez esetben az imputálandó folytonos változó tekintetében a nem adathiányos megfigyelésekre, mint célváltozó értékeire lineáris regressziós modell készül az adatbázis nem adathiányos egyéb változóival, mint prediktorokkal. Ezek után azokra az adathiányos helyekre, ahol a célváltozó értéke hiányzik, a regresszió segítségével becslést készítünk. A hagyományos statisztikai szoftverek nem tesznek különbséget valódi és az imputált értékek között, így a regresszióval pótolt értékek varianciája, a regressziós becslés logikájából következően kisebb lesz, mint a valóságban. Ennek orvoslására kerültek kifejlesztésre a *sztochasztikus regressziós imputálások*, melyeknél egy véletlen változót is adnak a regresszált értékhez, a standard hibák jobb becslése érdekében.

Hot deck imputáció

A hiányzó adatot tartalmazó megfigyeléshez, valamilyen hasonlósági mérték szerinti, leginkább hasonló, úgynevezett donor eset értékével történik a hiányos eset hiányzó érték pótlása. A hasonlóság mértékének megítélésére különböző módszerek használhatók (Andridge, R.R. & Little, R.J.A. 2010). A legtöbb esetben statisztikailag kedvező tulajdonságai és a továbbiakban ismertetendő módszerekhez képesti viszonylagos egyszerűsége (Roth, 1994) révén egyre növekvő népszerűségnek örvend az alkalmazók körében. A módszer hátránya, hogy nem könnyű feladat az esetek hasonlóságát definiálni különböző mérési szintű változók esetén. Továbbá a modellezőnek sokszor, alkalmas szoftver megoldás híján, saját programot kell készítenie a donor egységek kiválasztásához. A továbbfejlesztett modellek több hasonló esetet is keresnek, és azokból véletlenszerűen választják ki a donor megfigyelést, vagy folytonos ismérvek esetén az átlagukat használják az imputációhoz.

Hot deck imputációra példaként, egy elterjedt megoldást az úgynevezett hiányzás hajlamossági score-ok használatát említhetjük (Oravecz 2008). A módszer logisztikus regressziót alkalmaz, hogy az Y függő változóban a hiányzás / nemhiányzás valószínűségét becsülje az X_i változók segítségével. A megfigyelési egységek az így kapott hiányzás hajlamossági score-ok alapján képzett kvantilisekbe csoportosíthatók. A csoportokon belül a nem hiányos esetekből visszatevéses mintavétel segítségével lehet imputálni a hiányzó értékeket. Az eljárás minden hiányzó adatot tartalmazó változóra megismétlődik.

Maximum Likelihood imputáció

Nagy minták esetén a modellek torzítatlan paraméterbecslését eredményezi a maximum likelihood imputáció (ML impuatation) még MAR tipusú adathiány esetén is. Ilyenkor egy modell kerül meghatározásra a megfigyelt adatokra és a becsléseket a modell melletti posterior valószínűségekre, vagy likelihoodra alapozzák. Hogy a módszert jobban megérthessük, tegyük fel, hogy van egy modellünk az Y-ra, melynek eloszlását az $f(Y|\theta)$ sűrűségfüggvénnyel írhatjuk le, ahol θ ismeretlen paraméter (Oravecz,2008). Legyen $Y = (Y_{megfigyelt}, Y_{hiányzó})$, ekkor $f(Y|\theta) = f(Y_{megfigyelt}, Y_{hiányzó}|\theta)$ az $Y_{megfigyelt}$,és az $Y_{hiányzó}$ együttes eloszlását leíró sűrűségfüggvény, az $Y_{megfigyelt}$ peremeloszlása pedig :

$$f\left(Y_{megfigyelt}\,|\theta\right) = \int f\left(Y_{megfigyelt}\,, Y_{hiányzó}|\theta\right) dY_{hiányzó} \tag{4.3}$$

A likelihood MAR adathiány esetén:

$$L\left(\theta|Y_{megfigyelt}\,\right) = \int f\left(Y_{megfigyelt}\,, Y_{hiányzó}|\theta\right) dY_{hiányzó} \tag{4.4}$$

Az ML becslés ezek utána következő egyenlet megoldásával kapható:

$$D_\ell\left(\theta|Y_{megfigyelt}\,\right) = \frac{\partial \ln L\left(\theta|Y_{megfigyelt}\,\right)}{\partial\theta} = 0 \tag{4.5}$$

A módszer egyik hátránya, hogy ha nincs a fenti egyenletnek zárt alakú megoldása, akkor iteratív módszerek alkalmazására van szükség. A gyakorlatban további problémát okoz az is, hogy feltételezéssel kell élnünk a változók együttes valószínűség eloszlására vonatkozóan.

Várakozás maximalizáció

A várakozás maximalizáció (Expectation Maximization (EM))egy két lépésből álló iteratív módszer a maximum likelihood becslésre MAR típusú adathiány esetén. Az első lépésben kiszámításra kerül a teljes adatokat tartalmazó állományra a loglikelihood várható értéke (expectation lépés). A második lépésben a kapott várható érték behelyettesítésre kerül a hiányzó értékek helyére és maximalizálódik a likelihood függvény (maximalizáló lépés), mintha nem lett volna hiányzó adat, így új paraméterbecslések keletkeznek. Ez az iteratív kétlépéses eljárás mindaddig folytatódik, míg a paraméterbecslések konvergens változása egy előre megadott értéknél kisebb nem lesz. Adott érték esetén a konvergenciához annál több iteráció szükséges minél több a hiányzó adat (Bilmes, 1998)

Többszörös imputáció

Kedvező statisztikai tulajdonságai[17] okán az ML imputáció alternatívájaként tekinthetünk a többszörös imputációra ((Multiple Imputation (MI)) Rubin, 1987). Az előbbivel szemben nagy előnye, hogy gyakorlatilag bármilyen adatszerkezet és modelltípus esetén alkalmazható (Allison,2013). A többszörös imputáció a statisztikai bizonytalanságot hivatott megjeleníteni az adatpótlási mechanizmus során. A módszer alkalmazása során a hiányzó értékek helyére több lehetséges értéket is imputálnak, ezáltal párhuzamosan több hiánytalan adatbázist készítenek az eredeti adatbázisból. Az így létrejövő adatbázisokon a ugyanazt a modelltípust lefuttatva különböző paraméterbecslés és a hozzájuk tartozó standard hiba sorozatokat kapunk. Rubin (1987) megmutatta, hogy az így keletkező különböző becslések hogyan egyesíthetők két egyszerű szabály segítségével.

Tételezzük fel, hogy m db imputált adatbázist készítettünk. Az adatbázisonkénti modellfuttatásokból további számítások érdekében rögzítsük a becsült paraméterek és standard hibák értékét. Legyen a $\hat{\theta}_j$ becsülni kívánt paraméter a j-edik adatmátrixból

[17] Az ML imputációhoz hasonlóan a paraméterbecslések konzisztensek és aszimptotikusan normális eloszlásúak.MAR adathiány mechanizmus esetén is.

(j=1,2,...,m). U_j pedig legyen a $\hat{\theta}_j$ standard hibája. A származtatott becslést az egyedi becslések átlagaként kapjuk:

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j \tag{4.6}$$

A becslés standard hibájának kiszámításához először az imputáción belüli varianciát:

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j \tag{4.7}$$

majd az imputációk közötti varianciát kell kiszámolni:

$$\beta = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\theta}_j - \bar{\theta}\right)^2. \tag{4.8}$$

A teljes variancia:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)\beta. \tag{4.9}$$

Ennek alapján az együttes standard hiba pedig értéke $\sqrt{T}.A\theta = 0$ nullhipotézist a$t = \bar{\theta}/\sqrt{T}$próbafüggvénnyel tesztelhetjük, ami Student-féle t-eloszlást követ az következő szabadságfokkal:

$$df = (m-1)\left(1 + \frac{m\bar{U}}{(m+1)\beta}\right)^2. \tag{4.10}$$

A számolt paraméterre ezek után intervallumbecslés készíthető. A gyakorlati alkalmazhatóság szempontjából fontos kérdés, hogy a mintavételi bizonytalanság szimulálásához hány imputációra van szükség. Rubin (1987) szerint az m imputáción alapuló becslés relatív hatékonysága végtelen számú imputáció hatékonyságához képest megközelítően$\left(1 + \frac{\gamma}{m}\right)^{-1}$,ahol γ a hiányzó információk becsült aránya

$$\gamma = \frac{r + \frac{2}{df+3}}{r+1} \tag{4.11}$$

és

$$r = \frac{(1 + m^{-1})\beta}{\bar{U}} \tag{4.12}$$

a relatív variancianövekedés.

Az m és γ különböző értékei mellett elérhető relatív hatékonyságokat mutatja az alábbi táblázat(Oravecz, 2008):

3. Táblázat.Többszörös imputációval elérhető relatív hatékonyság (%)

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Megfigyelhető, hogy abban az esetben, amikor az adathiány mértéke nem túl magas, akkor már kevés számú imputáció is elégséges a statisztikai bizonytalanság megjelenítésére a modellben.

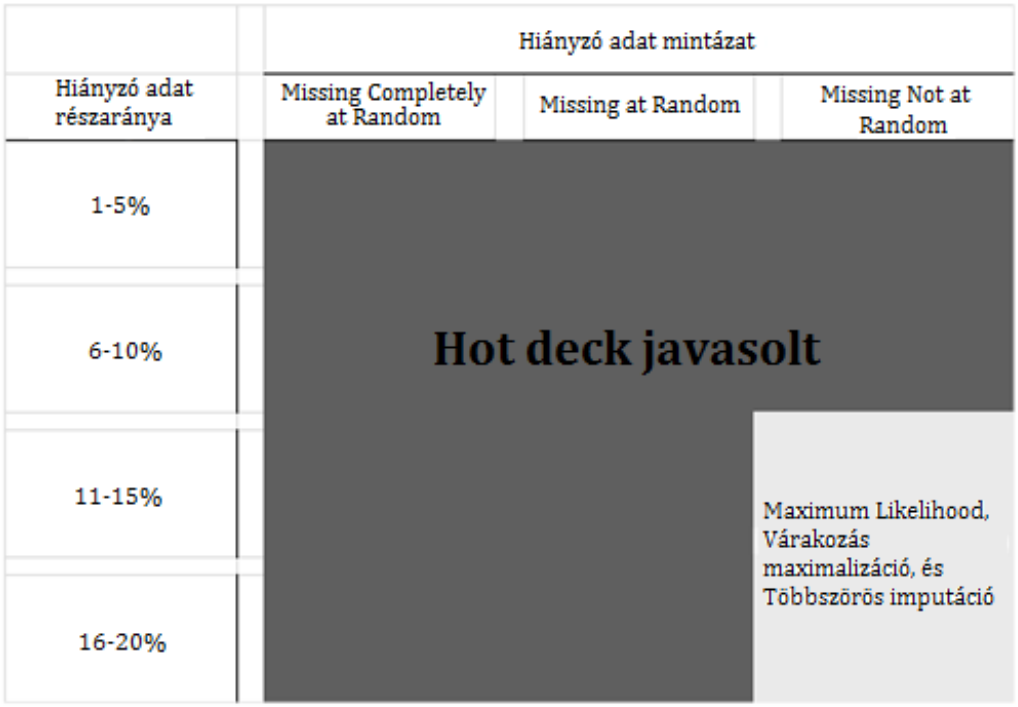
A módszer hátrányaként említhetjük, a számottevő adathiány esetén tapasztalható számításigényességét.

IV/5.Konklúzió

A hiányzó adatok kezelésének tudománya egy friss, manapság is dinamikusan fejlődő terület az alkalmazott statisztikán belül. A legjobb gyakorlatok, standardok kialakulása napjainkban is zajlik. A kereskedelmi szoftverekben található megoldások is legtöbbször gyártóról gyártóra változnak. Az egyszerű módszerek (listwise deletion, pairwise deletion,

egyszerű középértékkel történő imputációk) bár a legtöbb szoftverben megtalálhatók és alkalmazásuk is egyszerű, statisztikailag nem adnak kielégítő eredményt. A skála másik végén található szofisztikált megközelítések (maximum likelihood alapú módszerek, többszörös imputációk) statisztikailag ugyan nagyon kedvező tulajdonságokkal rendelkeznek, viszont idő és számításigényességük számos gyakorlati helyzetben megkérdőjelezik a használhatóságukat. Gyors elterjedésük további korlátját jelenti a gyakorlatban, hogy nem minden szoftver esetében állnak rendelkezésre, és viszonylagos újdonságuk és összetettségük révén a kutatók is fenntartásokkal közelítenek hozzájuk. Egyszerű kivitelezhetőségük és megfelelő statisztikai tulajdonságaik okán Roth (1994) a gyakorlati esetek többségénél a hot deck imputációs módszert javasolja alkalmazni a következő ábra dimenzió mentén:

6. ábra.Imputációs módszerek alkalmazása



V. Outlierek detektálása és kezelésük

Mivel a folytonos változók szélsőséges, kiugró értékei az úgynevezett outlierek általában rontják az alkalmazott modell illeszkedését, ezért a modellezés során külön gondot kell fordítani a szélsőséges, kiugró értékek felismerésére és megfelelő kezelésére. Különösen fontos ez a lineáris modellek esetén, ahol a kovariancia mátrix számítása során felhasznált folytonos változók átlagai jelentősen módosulhatnak a szélsőséges értékek következtében, nagy hatást gyakorolva ez által magának a mátrixnak egyes elemeire is. A következő példában a „KOR" folytonos életkor változóval szeretnénk prediktálni a műtét utána szövődményes állapot bekövetkezését. Az előrejelzéshez dichotóm logisztikus regressziós modellt alkalmazunk. Az adatbázis 350 beteget tartalmazott, melyből 98 esetben következett be szövődmény. A legfiatalabb beteg 15, a legidősebb 87 éves volt. A normális eloszláshoz hasonló eloszlás átlaga 54,86 mediánja 56 év. A „KOR" magyarázó változó felhasználásával futtatott logisztikus regresszió paraméterbecsléseit[18] foglalja össze az alábbi táblázat.

4. Táblázat. Paraméterbecslés.

Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	KOR	,032	,009	13,120	1	,000	1,032
	Constant	-2,735	,521	27,558	1	,000	,065

Látható, hogy a „KOR" szignifikáns magyarázó változó (*P=0,000*) a modellben. Következő lépésben szövődménymentes 77 éves beteget 577 és egy szövődményes 84 éves beteget 384 évesre „változtattunk". Ezáltal létrehoztunk két outlier megfigyelést. A regressziót újra futtatva a következő eredményeket kaptuk:

[18] A modellezéshez SPSS 20 programcsomagot használtuk.

5. Táblázat. Paraméterbecslés.Outlier

Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	KOR	,002	,003	,524	1	,469	1,002
	Constant	-1,072	,213	25,335	1	,000	,342

Látható, hogy nemcsak a paraméterek értéke változott meg, hanem a magyarázó változó szignifikancia értéke (*P=0,469*) is . A „KOR" változó semmilyen szokásos szinten nem tekinthető szignifikánsnak ebben a modellben. Példánk mutatja, hogy a szélsőséges megfigyelések megfelelő kezelése nélkül modelljeink illeszkedése, és ez által magyarázó ereje jelentősen romolhat.

A problémát alaposabban szemügyre véve természetesen megfontolás tárgyát képezi, hogy egyáltalán melyik megfigyelést tekinthetjük outliernek. Barnett és Lewis (1994) definíciója szerint ezek azok az esetek, melyek „inkonzisztensek" a többi megfigyelés eloszlásának tekintetében. A meghatározás problémája, hogy a kiugró értékek meghatározását alapvetően szubjektív értékítélettől teszi függővé, mely személyről személyre változhat.

Ezt kiküszöbölendő terjedt el az a megközelítés, hogy adatokról azt gondoljuk, hogy egy adott eloszlás (empirikus) által generálódnának és az outlierek esetében feltételezzük, hogy más eloszlásból származnak, mint az adatok fennmaradó része (Hawkins, 1980). A módszer gyakorlati alkalmazására az adatokhoz egy jól illeszkedő elméleti eloszlást próbálnak illeszteni, melynek segítségével tesztelhető (Grubb's és Dixon statisztikák, 1950), hogy egy-egy outlier jelölt megfigyelés milyen valószínűséggel származhat az adott elméleti eloszlásból. Csak érdekességképpen, ez a megközelítés a nagy népszerűségnek örvendő szabad forráskódú *R* programcsomagban is leprogramozásra került (van der Loo, 2010). A megközelítés kritikájaként fogalmazódik meg a módszer számításigényessége, valamint az, hogy sok esetben nem található jól illeszkedő elméleti eloszlás.

Egyszerű kivitelezhetőségük során nagy népszerűségnek örvendenek a gyakorlati modellezők körében az önkényes meghatározáson alapuló outlier definíciók. Ezek a módszerek közösek abban, hogy az adott változó tapasztalati eloszlása esetében valamilyen logika alapján önkényesen meghatároznak egy alsó és egy felső határértéket az eloszlásban, melyeknél kisebb illetve nagyobb értékeket már kiugró értéknek tekintenek. A kiugró értékeket ezek után úgy kezelik, hogy az alsó határértéknél kisebb értékeket az alsó, míg a

felső határértéknél nagyobb elemeket a felső határértékkel helyettesítik. Következőkben röviden áttekintünk néhány széles körben alkalmazott módszert ezek közül.

V/1.Winsorizálás

Winsorizálás (Vargha, 2008) során a nagyság szerint sorba rendezett változó valamelyik alsó, illetve felső percentilisét tekintjük kitüntetett határértéknek, melyeken túli értékek már outliernek minősülnek. Gyakori választás szokott lenni az első és a kilencvenkilencedik percentilis kijelölése erre a célra, de nemritkán előfordul az ötödik és a kilencvenötödik percentilis meghatározása is. Korábban említett módon, az így definiált kiugró értékeket a megfelelő határértékkel helyettesítik a további számolások végrehajtása érdekében.

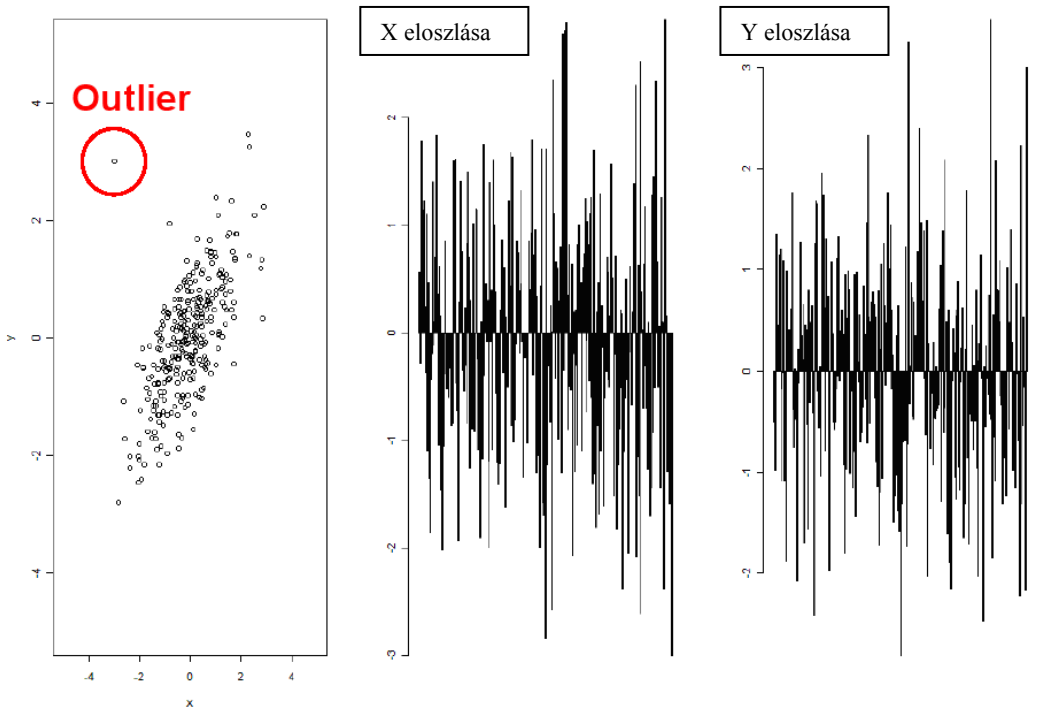
V/2.Szigma megközelítés

Ennél a módszernél kiugró értéknek azokat a megfigyeléseket tekintjük, melyek az eloszlás átlagától valahány szigma távolságnál messzebb vannak, ahol szigma az eloszlás empirikus szórása. A kiugró értékek helyettesítése ezek után ugyanúgy történik, ahogyan az a winsorizálás során bemutatásra került. A gyakorlatban igen elterjedt a három szórás körüli távolság meghatározása outlier határértékként.
Mind a winsorizálás, mind a szigma módszer egyváltozós módszer, azaz szélsőséges eseteket csak a saját eloszlásuk viszonylatában vizsgálja, így nem alkalmas arra, hogy segítségükkel megállapítsuk, hogy a több változó által kifeszített térben egy megfigyelés vajon kiugrónak tekinthető-e, avagy sem.

V/3.Mahalanobis távolságon alapuló megközelítés

A Mahalanobis távolságon alapuló módszer egy többváltozós megközelítés, és abból indul ki, hogy megfigyelés akkor is lehet többdimenziós értelemben kiugró, az eloszlás jellegétől alapvetően eltérő, ha egyébként változónként külön-külön vizsgálva nem tekinthető outliernek. Az ilyen sajátosságot mutató esetek az egyváltozós vizsgálat során rejtve maradnak, detektálásuk ezen a módon nem lehetséges. Ilyen esetre mutat példát két változó esetén az alábbi ábra.

7. ábra. Kétdimenziós outlier



Az ábrán jól látszik, hogy x és y között erős pozitív lineáris korrelációval jellemezhető kapcsolat van. Két megfigyelés kölcsönös helyzetének jellemzésekor a Mahalanobis távolság (***MD***), mint metrika figyelembe veszi a változók között fennálló lineáris kapcsolatrendszert is, a következő módon:

$$MD(x) = \sqrt{(x-\mu)^T \sum^{-1} (x-\mu)} \qquad (5.1)$$

Ahol, $x$az adott megfigyelés vektora $\sum^{-1}$ a kovarianca mátrix inverze, μpedig az adatmátrix átlagvektora, centroidja. A képlet tehát egy megfigyelésnek az adathalmaz „középpontjához” viszonyított elhelyezkedését jellemzi.

Legyen például x_1 és $x_2$változók alkotta adatmátrixban

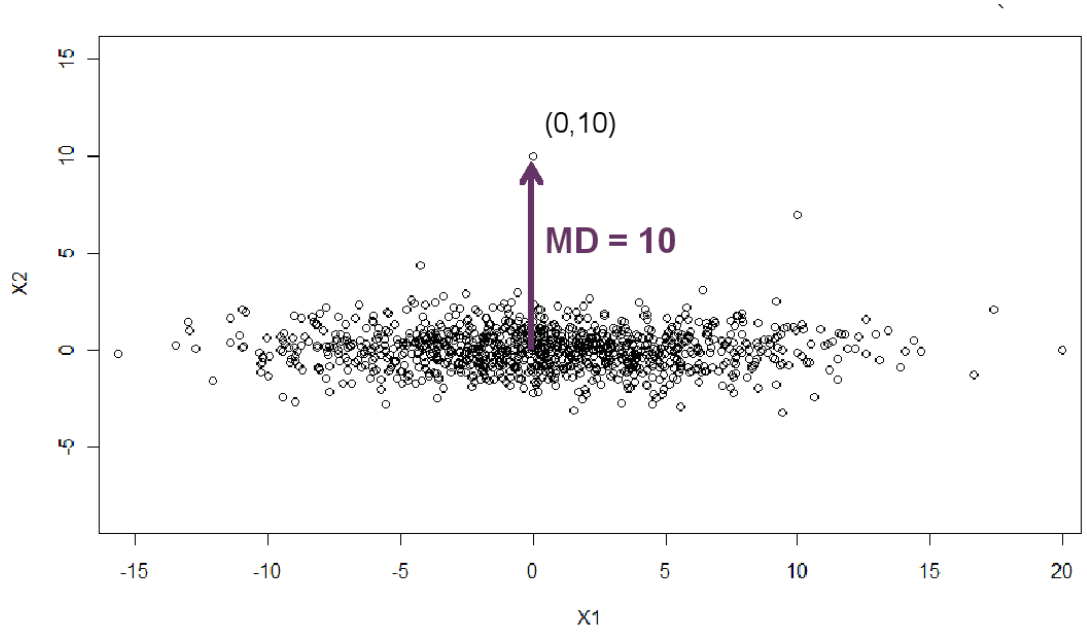
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5.2}$$

és

$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix} \tag{5.3}$$

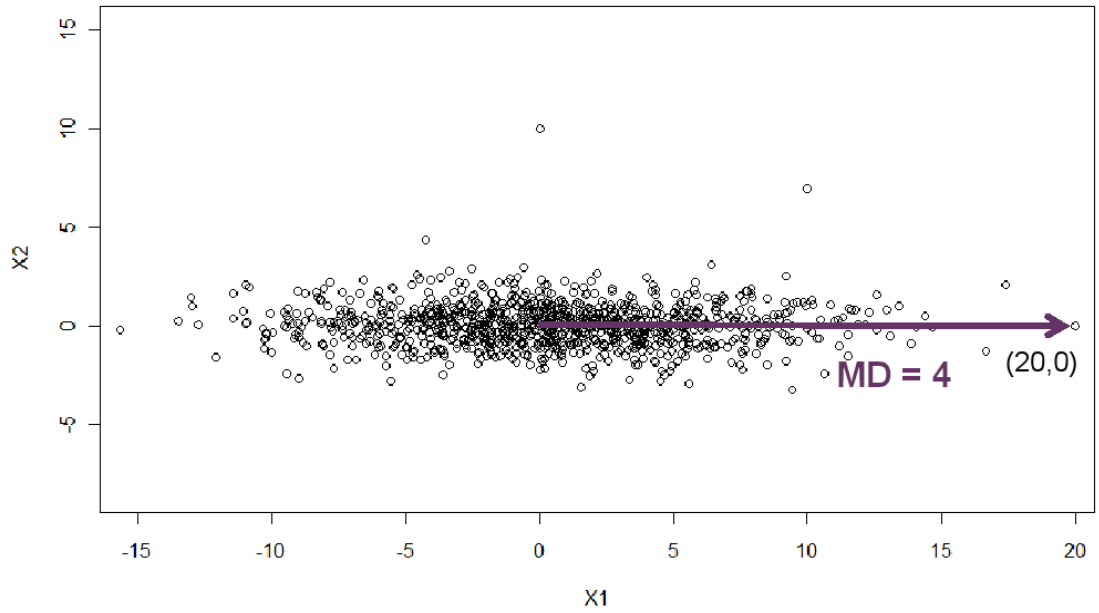
ekkor két kiválasztott pontra kiszámított távolságok láthatók az alábbi ábrákon:

8. ábra. Mahalanobis távolság 10



valamint

9. ábra. Mahalanobis távolság 4



Látható, hogy a két megfigyelés mindegyike outliernek tekinthető egy-egy változóban. A második esetben azonban, bár az Euklideszi távolság jelentősen nagyobb az adatállomány centroidjától, a Mahalanobis távolság többszörösen kisebb, ami annak köszönhető, hogy a megfigyelés a megfigyelések pontfelhőjének tendenciájához (a regressziós egyeneshez) közel esik. A többdimenziós outlierek figyelembevétele ezek után úgy történik, hogy minden megfigyelésre meghatározásra kerül a Mahalanobis távolság a fenti módon. A távolságok kiszámítását követően a távolságok reciprokjaival, mint súlyokkal kerül futtatásra a regresszió. Így biztosítható, hogy a magasabb MD értékű megfigyelések kisebb szerepet kapjanak a paraméterbecslések során.

V/5.Konklúzió

Az alapvető módszerek megismerése után felvetődik a kérdés, hogy melyik módszert érdemes használni a gyakorlati munka során. Tiwari, Mehta, és Jain 2007-ben publikált empirikus kutatása alapján a logisztikus regresszió esetében a nagyon egyszerűen megvalósítható szigma megközelítés alapján kezelt kiugró értékekkel futtatott regresszió bizonyult a legeredményesebbnek. Az eredmények mindazonáltal bizonyos óvatossággal értékelendők, tekintve, hogy a szerzők csak egy adott minta esetében vetették össze a különböző outlier kezelési módszereket.

VI. Változók transzformációja

VI/1.Előzetes változószelekció

Amennyiben az adatbázisra a választott modell tekintetében túl alacsony EPV (Events Per Variable) jellemző[19], szükségessé válik a változószám csökkentése. A modellezés során azok a változók a legkevésbé értékesek, melyeknél a hordozott információtartalom nem vagy csak csekély hozzájárulással bír az előrejelezni kívánt változó vonatkozásában. A lehetséges magyarázó változók és a célváltozó közötti egyváltozós kapcsolat erősségét alkalmasan megválasztott kapcsolat szorossági mérőszámmal tudjuk jellemezni. Közkedvelt és minden statisztikai programcsomagban megtalálható a khi-négyzet statisztika, melynek segítségével kétkategóriás mérési szintű változó közötti kapcsolat erősségét jellemezhetjük. Kategóriás mérési szintű változók esetén ezek után a célváltozó tekintetében minden kategóriás magyarázó változó esetében kiszámoljuk a khi-négyzet statisztikákhoz tartozó empirikus szignifikancia (*p*-érték) értékeket. Folytonos mérési szintű változók esetén a khi-négyzet statisztika használatának feltétele az, hogy először a folytonos változót célszerűen megválasztott számú osztóponttal kategória változóvá alakítjuk.[20]Az így kategorizált változókon a khi-négyzet statisztikákat és a hozzájuk tartozó empirikus szignifikanciákat az előzőekhez hasonlóan számoljuk. Az eljárás végére minden változóra rendelkezni fogunk empirikus szignifikancia értékekkel. A változókat ezen értékek mentén sorba rendezve az alacsonyabb p-értékek jelzik a szorosabb, míg a magasabb értékek a kevésbé szoros kapcsolatokat. A szűrést ezek után úgy hajtjuk végre, hogy a legmagasabb p-értékkel rendelkező változók közül elhagyunk annyit, amennyi a minimális EPV arány eléréséhez szükséges. A gyakorlatban előfordulhat, hogy az adatbázis nagyon sok folytonos mérési szintű változót tartalmaz. Megfelelő szoftver támogatás hiányában a folytonos változók kategorizálása rendkívül időigényes lehet. Ilyenkor a folytonos magyarázó változó és a bináris célváltozó kapcsolatának egyfajta jellemzésére használhatjuk a minden programcsomagban megtalálható F-statisztikát. A tesztelendő hipotézis ebben az esetben az, hogy a célváltozó két kategóriájában a folytonos változó

[19] Enter vagy backward változószelekciós módszer használata esetén fordulhat elő leginkább ez a helyzet, mert ilyenkor az összes változó egyszerre kerül be a modellbe.

[20] Szerencsére a módszer szoftver támogatása egyes statisztikai/adatbányászati programcsomagban megtalálható Az eljárás közben az eredeti folytonos változót célszerű megtartani, annak érdekében, hogy amennyiben szükséges a változó eredeti mérési szintjével tudjunk továbbdolgozni.

eloszlásának részátlagai tekinthetők-e egyenlőnek. Az empirikus szignifikanciák itt is számolhatók, ennek alapján a változók rendezésére és szűrésére a korábban ismertetett módon lehetőség nyílik. A módszer árnyoldala az, hogy nem kellően pontos, azaz egyes esetekben előfordulhat, hogy a részátlagok között nincs szignifikáns különbség, míg a folytonos változó kategorizált párja esetében a khi-négyzet statisztika alapján kapcsolat mutatkozna célváltozóval, vagy ellenkezőleg részátlagok szignifikáns különbsége esetén a khi-négyzet alapján nem mutatkozik számottevő kapcsolat a két változó között[21]. Emiatt csak indokolt esetben javasolt a módszer használata.

A gyakorlatban, a határidők nyomása miatt idő szűkében dolgozó modellezők esetében, szokott előfordulni előzetes, jellemzően stepwise forward logisztikus regresszió[22] futtatása a változók előszűrése céljából. A regresszió futtatása után csak a modellbe bekerült változókkal folytatnak további modellezést, a többi változót elhagyják a modellezési adatbázisból. A módszer előnye a gyors és egyszerű kivitelezhetőség, valamint az, hogy a bennmaradó változók a többváltozós kapcsolatrendszerük viszonylatában bizonyulnak szignifikánsnak. A módszer hátránya, hogy a modellből nagy valószínűséggel kiesnek azok, az egyébként esetleg jelentős prediktív erővel bíró folytonos változók, melyek a célváltozóval nem monoton kapcsolatban állnak. Ennek magyarázatára a folytonos változók kapcsolata a célváltozóval c. fejezetben a későbbiekben még sor kerül.

VI/2.Előszűrés problémái – változók értékelésének egyváltozós és többváltozós megközelítése

Az előszűrés fent leírt khi négyzet alapú eljárása egyváltozós (univariate) statisztikai technikán alapszik. A módszer célja, hogy a változók egyedi előrejelzési ereje alapján a változókat rangsorolni tudjuk. Az eljárás eredményeképpen eltávolítunk az adatbázisból a célváltozóval látszólag gyenge vagy kapcsolatban nem lévő magyarázó változókat. A problémát az okozza, hogy az egy változóban független változó többváltozós környezetben akár jelentős parciális hatással bírhat a függővváltozó értékére, így elhagyása a modellből a predikciós erő csökkenését okozhatja. Tegyük fel például, hogy a lakásárak alakulására akarunk egy regressziós modellt készíteni. A mintát a budapesti Rózsadomb és a nyolcadik

[21] Az F-próba alapfeltétele a csoportonkénti eloszlások normális mivolta és a szórások egyezősége. A valóságban ezek a feltételek leggyakrabban sérülnek.

[22] A több száz változóval rendelkező adatbázisok esetén a backward és a forward változószelekciós algoritmus futási időigénye között jelentős eltérés szokott mutatkozni.

kerület megfigyelései alkotják. Az elemzés során azt találjuk, hogy az egyváltozós elemzés alapján a lakásárak és a lakások alapterülete között nincs statisztikailag szignifikáns kapcsolat. A többváltozós regresszió számítás során azonban a lakás nagysága szignifikáns magyarázó változónak bizonyul. A jelenség magyarázata alapvetően az, hogy a két kerületben a lakások négyzetméter ára nagyban különbözik, a Rózsadombon a kis lakásnak is lehet olyan magas ára, mint a nyolcadik kerületben egy nagy lakásnak. Így látszólag a lakás nagysága nem hat az árra, azonban ha a területi hovatartozást is bekapcsoljuk a modellbe a nagyság parciális hatása már jelentős lesz. A lakásokat a Rózsadombon és a nyolcadik kerületben külön-külön vizsgálva az ár és a nagyság között már van kapcsolat. A lakásnagyság változó előzetes kiszűrése az adatmintából tehát a modell előrejelző képességének romlását okozta volna. Hogy milyen jelentős következményei lehetnek, az ilyen természetű egyváltozós tévkövetkeztetéseknek álljon itt egy saját kutatási tapasztalat. Pár évvel ezelőtt egy hazai egészségügyi intézet műtéti részlegén a műtét utáni szövődmények kialakulásának előrejelzése volt a feladatunk. A műtéti eseteket két csoportba soroltuk annak megfelelően, hogy a műtétet követően kialakult-e az illető betegnél szövődmény, vagy sem. Az orvosi kutatásoknál megszokott és elvárt módon elvégeztük a szövődmény bekövetkezése és a lehetséges magyarázó faktorok egyváltozós kapcsolatának elemzését. Meglepő módon a dohányzási szokásokat leíró változónál azt tapasztaltuk, hogy minél több cigarettát szív el valaki naponta annál kisebb a szövődmény bekövetkezésének esélye. Az egyváltozós elemzés alapján úgy tűnt tehát, hogy a dohányzás védő hatású a műtét utáni szövődmény bekövetkezése tekintetében! A józan észnek ellentmondó eredmény azonnal érthetővé vált akkor, amikor a dohányzási szokásokat a többi változóval együtt vizsgáltuk. Azt tapasztaltuk, hogy az adatmintában a fiatalok körében jóval magasabb volt a dohányosok aránya, mint az idősebbek esetében, így hipotézisünk szerint a magasabb cigarettaszám valójában az alacsonyabb életkor információját hordozta magában és a dohányzás parciális hatása ténylegesen negatív a szövődményráta alakulására nézve. A többváltozós elemzés igazolta ezt a vélekedést azáltal, hogy dohányzás és az életkor növekedésének parciális hatása is negatívnak bizonyult.

VI/3.Folytonos változók kapcsolata a célváltozóval

A folytonos változók regressziós/prediktív modellben betöltendő szerepére hatást gyakorol, hogy van-e és ha van, milyen jellegű a kapcsolata a célváltozóval. A kapcsolat jellemzése céljából folytonos változó kategorizált párjának kategóriáira vonatkozóan bevezetjük a számos programcsomagban elterjedt WOE (Weight Of Evidence) fogalmát. A könnyebb megértés kedvéért tegyük fel, hogy a modellezendő célváltozónk bináris, és egy hitelintézet ügyfeleinek fizetőképességét kódoljuk vele. Legyen a célváltozó értéke a „jó" amennyiben az adott ügyfél rendben megfizette a tartozását és legyen a változó értéke „rossz" ha a hitel visszafizetése nem történt meg. A kategorizált folytonos változó c-edik kategóriájára a WOE értékét esetünkben a következő módon számítjuk:

$$WOE_c = \ln\left(\frac{p_c^{jó}}{p_c^{rossz}}\right) = \ln\frac{1 - p_c^{rossz}}{p_c^{rossz}} \tag{6.1}$$

ahol

$$p_c^{jó} = \frac{N(jó)_c}{N(jó)} \tag{6.2}$$

$$p_c^{rossz} = \frac{N(rossz)_c}{N(rossz)} \tag{6.3}$$

Amennyiben$p_c^{jó} = 0$, akkor az adott kategóriára

$$WOE_c = 1/N(jó) \tag{6.4}$$

és

$$p_c^{rossz} = 0 \tag{6.5}$$

esetén

$$WOE_c = 1/N(rossz) \tag{6.6}$$

Intuitíven a WOE mutató a c.-edik kategóriába eső jól és rosszul teljesítő ügyfelek relatív gyakoriságainak arányára enged következtetni. Ha például az ügyfelek fizetési hajlandósága az életkor növekedésével egyre jobb lesz, akkor az életkor folytonos változó kategorizált párjának növekvő életkor kategóriái mentén a WOE értéke kategóriáról kategóriára is növekszik. A képletet vizsgálva az is világos, hogy egy adott kategóriára jellemző WOE csak akkor lehet nulla, ha az adott kategóriában a jó és rossz ügyfelek százalékos aránya megegyezik. A kategóriánkénti WOE értékeket alkalmasan összegezve kapjuk az egész változóra jellemző Information Value (IV) értéket:

$$Information\ Value = \sum \left(p_c^{jó} - p_c^{rossz}\right) \times WOE_c \qquad (6.7)$$

Az IV mutató felfogható egyfajta kapcsolat szorossági mérőszámként is, magasabb értékei a célváltozóval való erősebb kapcsolatot jelentik[23].
Adott változó esetében a kategóriánkénti WOE értékek vizsgálatával képet alkothatunk a folytonos változó és a célváltozó kapcsolatáról.

VI/4.Folytonos változók kategorizálása

Mind a lineáris, mind a logisztikus regressziós modell esetén, az alkalmazott függvénytípus következtében, csak a célváltozó tekintetében monoton kapcsolatot mutató folytonos változók esetében várhatunk megfelelő illeszkedést. Ezért a nem monoton kapcsolatot mutató változók[24] esetén a folytonos változót kategorizált alakjában javasolt szerepeltetni a modellben a prediktív erő növelése céljából. Kevésbé magától értetődő, hogy nemegyszer monoton kapcsolat esetén is előállhat olyan eset, hogy a folytonos változó nem szignifikáns, míg a kategorizált párja szignifikáns magyarázó változónak bizonyul modellünkben. Erre a helyzetre mutat példát a következő eset, amikor egy egészségügyi adatbázisban[25] az életkor (KOR) és a testtömegindex (BMI) folytonos változókkal szeretnénk előrejelezni a műtét utáni szövődmények (TARGSSI) előfordulását.

[23] A változók előszűrésénél a khi-négyzet statisztika alternatívájaként is használható.
[24] Tegyük fel, hogy a fizetéseket akarjuk regresszálni az életkor segítségével. A fizetések jellemzően az életkor előrehaladásával eleinte növekszenek, majd egy idő után csökkeni kezdenek. A kapcsolat fordított U alakú, folytonos formájában szerepeltetve a modellben, jó eséllyel nem szignifikáns változóként fog szerepelni.
[25] Az adatbázis 350 beteg adatait tartalmazta, akik közül 98 esetben fordult elő műtét utáni szövődmény. Az outputok SPSS 20 programcsomaggal készültek.

Az alábbiakban láthatjuk a két változót egyszerre a modellbe bevonó (enter) logisztikus regressziós modell paraméterbecslésével kapcsolatos statisztikákat feltüntető táblázatot.

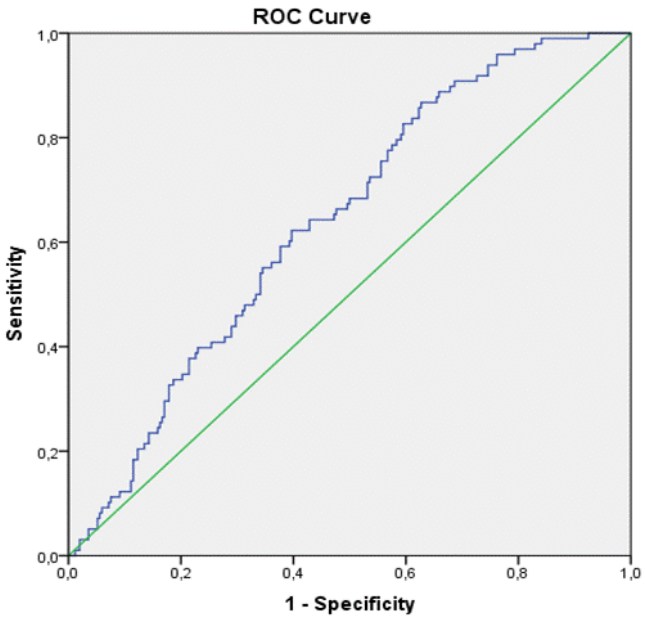
6. Táblázat. Paraméterbecslés (KOR, BMI)

Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	BMI	,010	,009	1,187	1	,276	1,010
	KOR	,032	,009	13,088	1	,000	1,032
	Constant	-3,024	,585	26,709	1	,000	,049

Látható, hogy a két változó közül csak a KOR változó szignifikáns a szokásos szinteken. Ennek megfelelően a modell illeszkedése is nagyon gyenge (GINI=28%) az alábbi ROC-görbének megfelelően:

10. ábra. ROC görbe (KOR, BMI)



Nézzük meg, mi történik, ha a testtömegindex folytonos változó helyett a kategorizált párját használjuk a modellezéshez. A kategóriahatárokat, az új változó (BMIKAT2) kódolását és az egyes kategóriákhoz tartozó szövődményrátákat (kategórián belüli szövődményes esetek aránya a kategória összes esetéhez) és WOE értékeket láthatjuk az alábbi táblázatban.

7. Táblázat. BMI kategorizálása

BMI	*0-28*	*28-34*	*34-*
BMIKAT2	*1*	*2*	*3*
SZÖVŐDMÉNYRÁTA	*12,70%*	*31,40%*	*44,60%*
WOE	*1,927748*	*0,781485*	*0,216846*

Látható, hogy kategóriák mentén monoton csökken a WOE értéke, tehát a BMI folytonos változó a célváltozóval való kapcsolata monoton, azaz a magasabb BMI értékek nagyobb szövődményrátákat vonzanak. A változó ennek ellenére nem bizonyult szignifikánsnak eredeti modellünkben. A regressziót újra futtatva az életkor folytonos és a testtömegindex kategorizált változóval a következő eredményeket kapjuk.

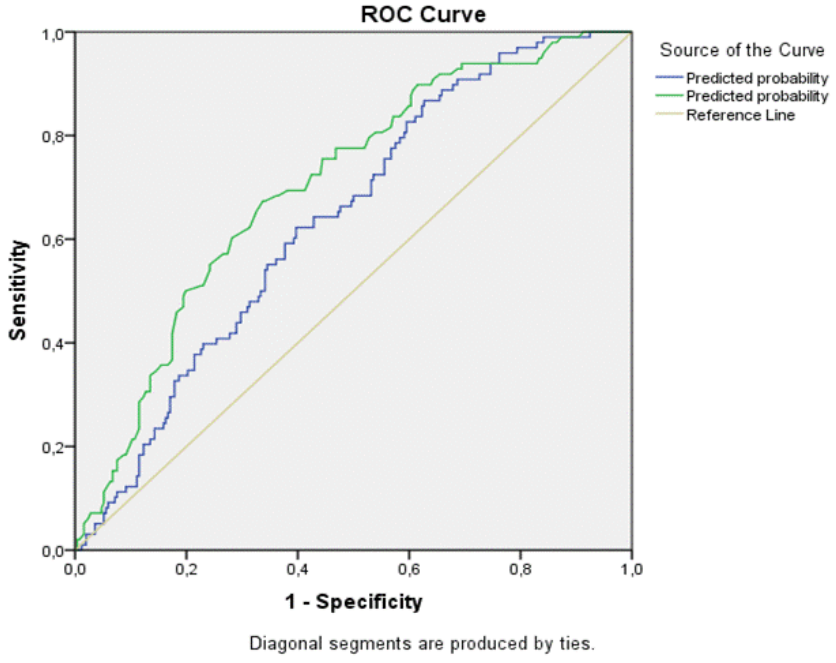
8. Táblázat. Paraméterbecslés(KOR, BMIKAT2)

Variables in the Equation

		β	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(β)*
Step 1	*KOR*	*,034*	*,009*	*13,767*	*1*	*,000*	*1,035*
	BMIKAT2	*,879*	*,189*	*21,561*	*1*	*,000*	*2,409*
	Constant	*-4,615*	*,711*	*42,113*	*1*	*,000*	*,010*

A testömegindex itt már szignifikáns és ennek megfelelően alakul a modell illeszkedése is, amely most már sokkal jobb (GINI=40%), mint az eredeti modellünk esetében. Az alábbi ábrán a két modellhez tartozó ROC-görbéket egyszerre tanulmányozhatjuk.

11. ábra. ROC görbe (KOR, BMIKAT2)



VI/5.Optimális kategorizálás algoritmusa

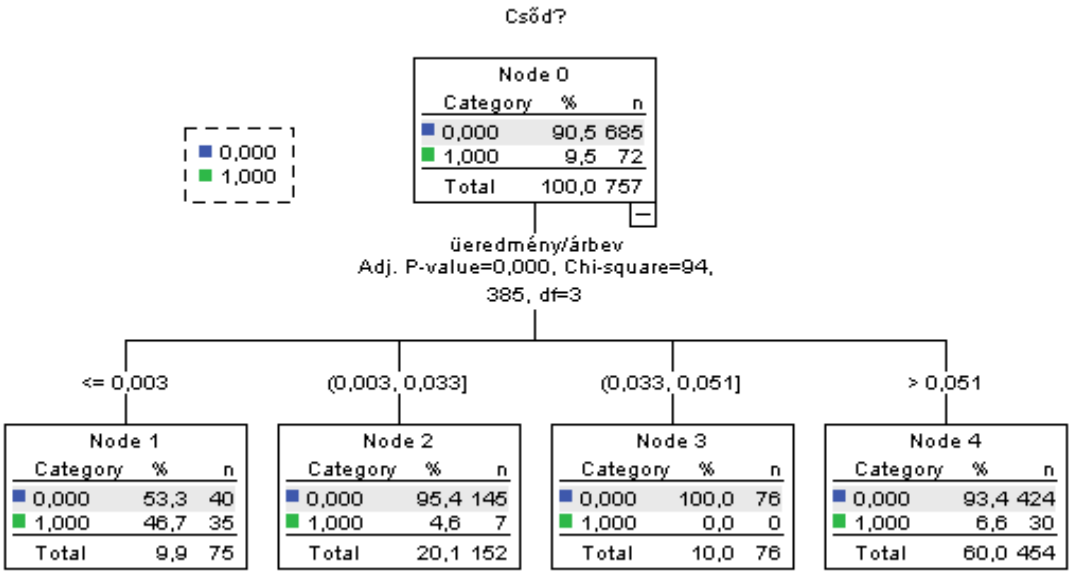
Miután megállapítottuk, hogy mely folytonos változó esetében szükséges a kategorizált párral történő helyettesítés, felmerül a kérdés, hogy hány kategóriával rendelkezzen az új változó, illetve hol legyenek a kategória határok[26]. Mivel a kategóriák száma és elhelyezkedése hatást gyakorol a végső változó predikciós erejére a modellben, nem mindegy, hogy a folytonos skála felosztását milyen módon hajtjuk végre. Az optimális felosztás megtalálására több algoritmust is használhatunk. Közös vonásuk, hogy a célváltozó tekintetében a felosztást úgy hajtják végre, hogy a létrejövő kategóriákon belüli homogenitás és a kategóriák közötti heterogenitás a legnagyobb legyen. A következőkben elérhetősége és egyszerűsége okán, egy közismert algoritmust alkalmazunk kategóriaegyesítési céllal[27]. Az algoritmus, egy a döntési fák családjába tartozó rekurzív klasszifikáló eljárás, az úgynevezett CHAID (Chi-squared Automatic Interaction Detector) algoritmus (Hámori, 2001) részalgoritmusaként alkalmazzuk. Az algoritmus célja, hogy a *K* különböző kategóriával rendelkező változó[28] esetében összevonásra kerüljenek azok a kategóriák, melyek legkevésbé különböznek egymástól az m különféle kategóriával rendelkező célváltozó tekintetében[29]. Ehhez X_i kategorizált folytonos változó kategóriái közül az összes lehetséges módon kiválaszt kettőt. Amennyiben a vizsgált magyarázó változó K különböző kategóriával rendelkezik, a kiválasztás $K^* (K-1) / 2$ féleképpen történhet. Ezt követően $K^*(K-1)/2$ különböző, (2 x m) méretű, kontingenciatáblára. Pearson féle khi-négyzet teszt segítségével kiszámolja, hogy milyen „*p*" szignifikancia szinten tekinthetők X_i kiválasztott kategória párjai és Y célváltozó kategóriái függetlennek egymástól. A következő lépésben kiválasztásra kerül az a kontingenciatábla, mely a legmagasabb „*p*" értékkel rendelkezik. Ezt az értéket az eljárás összeveti egy, a modellkészítő által előre lerögzített, $\alpha_{egyesítés}$ küszöbértékkel (a programcsomagok általában a szokásos 5%-os szignifikancia szintet szokták felkínálni alapértelmezésként).Amennyiben $p>\alpha_{egyesítés}$ a kontingenciatáblázat $X_i$kategóriapárja egy új önálló kategóriába kerül egyesítésre. Ebben az esetben X_i eredeti kategóriáinak száma eggyel csökkent, és az

[26] A szakzsargonban a kategorizálási eljárást angolul „data-binning" illetik. Az elnevezés után a hazai adatbányász szóhasználatban közkedvelt a „változó binnelése" fogalom használata.
[27] Az algoritmus kategóriaváltozók esetén is alkalmas a változó meglévő kategóriáinak optimáló kritériumok mellett történő összevonására, ezáltal csökkentve a kategóriák számát.
[28] Folytonos változók esetén az algoritmus alapbeállításként a K=10 értéket ajánlja fel. Ebben az esetben a változó decilisei jelennek meg, mint kategóriák.
[29] Logisztikus regresszió esetében m=2

algoritmus újból indul az elejétől, azaz az „új” kategóriapárok kiválasztásától (amelyek között nyílván lehetnek olyanok is, melyek az előző ciklusban is kiválasztásra kerültek), az azokhoz rendelt kontingenciatáblákhoz tartozó „*p*” értékek kiszámolásáig.

A kategóriák összevonásának ciklusa mindaddig folytatódik, míg a legmagasabb *„p”* értékkel rendelkező kontingenciatáblára igaza nem lesz a $p<\alpha_{egyesítés}$ feltétel. Ekkora vizsgált magyarázó változó *(X_i)* esetében a ciklus leáll, és az algoritmus a következő lépésben most már *X* teljes, lehetséges összevonások utáni**,** új kategória-struktúrájára kiszámolja a *„p”*értékét. Az így létrehozott új változó most már alkalmas arra, hogy a prediktív modell lehetséges magyarázó változójaként a modellépítés során felhasználásra kerüljön. A következő ábra mutat példát egy konkrét folytonos változó esetében az algoritmus végeredményére[30].

12. ábra.CHAID alapú kategorizálás



A kategorizálandó pénzügyi mutató az *üzemieredmény/árbevétel*. Az ábra téglalapjaiban láthatók a célváltozó (Csőd) lehetséges értékei (0,1) szerinti megoszlások a pénzügyi

[30] Az ábra az SPSS answer tree CHAID programmoduljának segítségével készült.

mutató kategóriái mentén[31]. A legfelső téglalapban (a fa csúcsán) látható, hogy a kiindulási adatbázis 685 fizetőképes és 72 csődös vállalatot tartalmazott. Az ábráról leolvashatók a kategóriahatárok, melyek rendre a következők:

- 0,003 az első és második kategória esetében
- 0,033 a második és harmadik kategória esetében
- 0,051 a harmadik és negyedik kategória esetében

Most nézzük meg, mi történik akkor, ha más kategorizálási logika mentén alakítjuk ki a kategóriahatárokat. Induljunk ki a folytonos alapváltozó eloszlását leíró fontosabb statisztikákból. A következő táblázat tartalmazza az *üzemieredmény/árbevétel* mutatót jellemző leíró statisztikákat:

[31] A felső téglalapban a teljes adatbázis megoszlását mutatja a „Csőd" változó kategóriái mentén.

9. Táblázat. Üzemi eredmény/árbevétel mutató statisztikái

Statistics

N	**Valid**	***757***
	Missing	***0***
Mean		***,0538***
Median		***,0645***
Mode		***,05***
Std. Deviation		***,40277***
Range		***9,17***
Minimum		***-7,58***
Maximum		***1,59***
Percentiles	**10**	***,0031***
	20	***,0196***
	25	***,0252***
	30	***,0328***
	40	***,0508***
	50	***,0645***
	60	***,0809***
	70	***,1044***
	75	***,1195***
	80	***,1378***
	90	***,1961***

Az alapstatisztikák felhasználásával definiáljunk további négyféle kategorizálási módszert a következő módon:

1. A három kvartilis segítségével meghatározott négykategóriás változó, melyet nevezzünk „Q123KAT” módon
2. Az eloszlást önkényesen négy részre felosztó pontok segítségével meghatározott „SCALEKAT” változó
3. A medián által meghatározott bináris változó, melyet jelöljük „MEDIÁNKAT” jelöléssel
4. Az átlag (mean) segítségével hozzuk létre a „MEANKAT” bináris kategóriaváltozót

Az CHAID-alapú optimális kategorizálás eredményeképpen létrejött változót nevezzük el „CHAIDKAT” módon, melynél a kategóriahatárok rendre a már ismert 0,003; 0,033 és 0,051.

VI/6.Alternatív kategorizálások vizsgálata

Vizsgálatunk célja az alternatív kategorizálási módszerek által előállított változók prediktív erejének értékelése az optimális kategorizálási módszer eredményeképpen előálló „CHAIDKAT” változóval szemben. A feladat végrehajtására egyváltozós logisztikus regressziókat futtatunk az egyes változókon, majd ROC görbével és GINI koefficienssel értékeljük az illeszkedést.

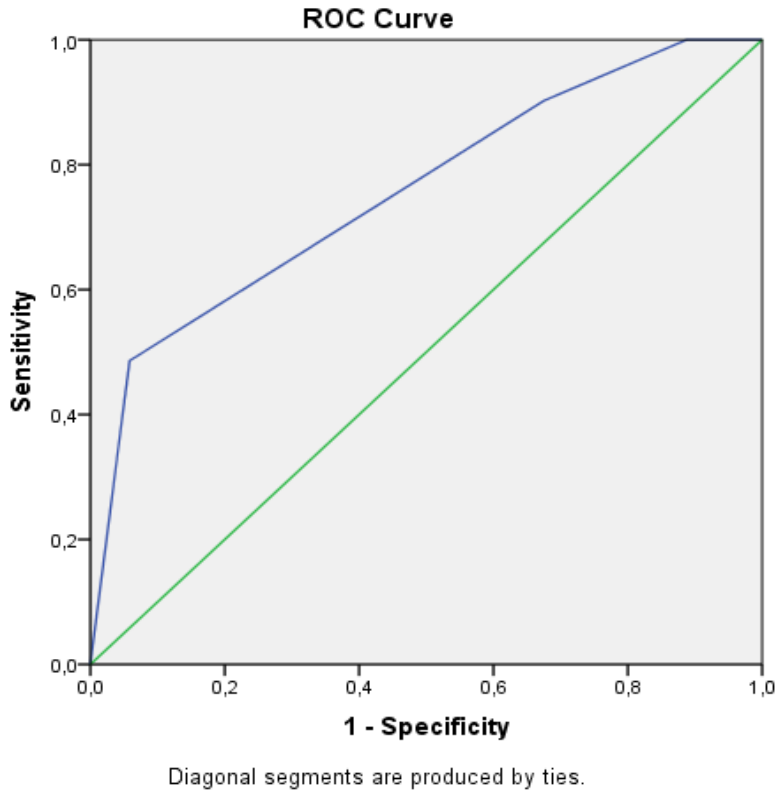
A „CHAIDKAT” változó esetében a regresszió paraméterbecslésének eredményei és a ROC görbe látható az alábbi táblázatban és ábrán:

10. Táblázat. A „CHAIDKAT” változó paraméterbecslése.

Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	chaidkat			82,001	3	,000	
	chaidkat(1)	2,513	,299	70,723	1	,000	12,338
	chaidkat(2)	-,385	,431	,798	1	,372	,681
	chaidkat(3)	-18,557	4580,414	,000	1	,997	,000
	Constant	-2,646	,189	196,155	1	,000	,071

13. ábra. A „CHAIDKAT” változó ROC görbéje



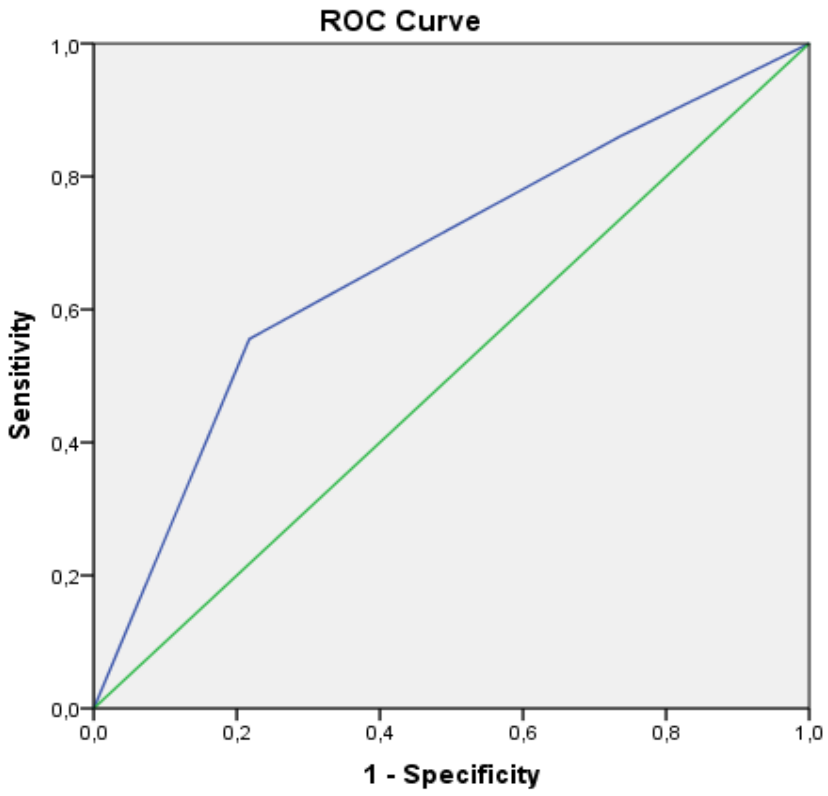
A GINI koefficiens értéke 51,4%.A „Q123KAT” esetében a logit regressziós eredmények a következőképpen alakulnak.

11. Táblázat. A „Q123KAT” változó paraméterbecslése

Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	Q123kat			34,883	3	,000	
	Q123kat(1)	1,463	,358	16,691	1	,000	4,320
	Q123kat(2)	-,011	,439	,001	1	,980	,989
	Q123kat(3)	-,112	,450	,062	1	,803	,894
	Constant	-2,778	,311	79,938	1	,000	,062

14. ábra. A „Q123KAT” változó ROC görbéje



A GINI mutató ebben az esetben már kisebb, mint az előző esetben, értéke 34,7%.

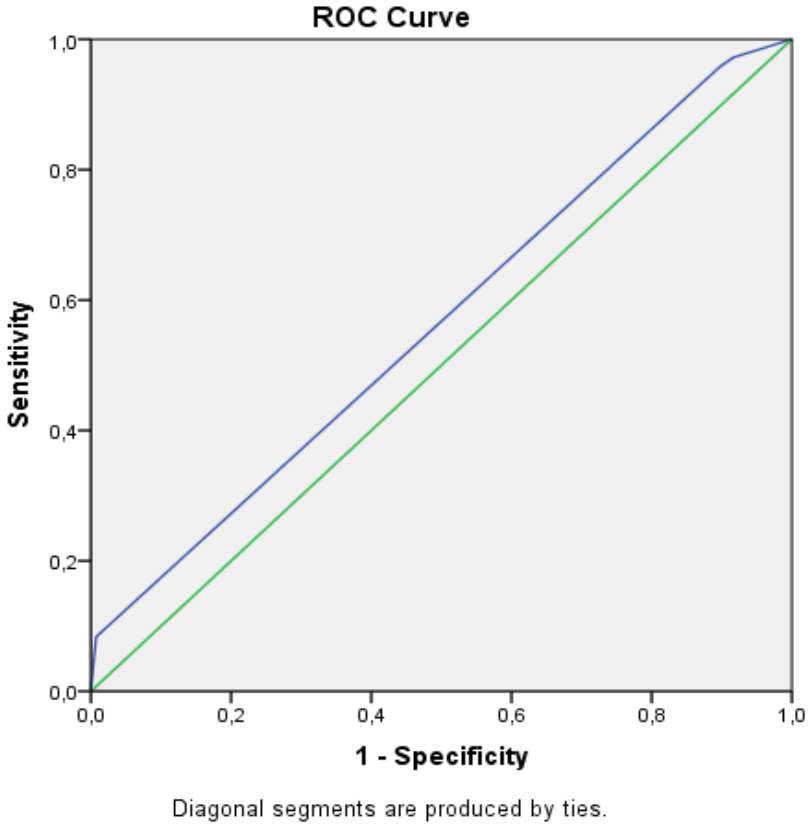
Az önkényes kategorizálás alapján létrejövő „SCALEKAT” változó esetében az eredmények a következő módon alakulnak.

12. Táblázat. A „SCALEKAT” változó paraméterbecslése

Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	scalekat			18,439	3	,000	
	scalekat(1)	2,747	1,201	5,228	1	,022	15,600
	scalekat(2)	,295	1,046	,079	1	,778	1,343
	scalekat(3)	-,785	1,263	,386	1	,534	,456
	Constant	-2,565	1,038	6,109	1	,013	,077

15. ábra. A „SCALEKAT” változó ROC görbéje



Az illeszkedés nagyon gyenge, ennek megfelelően a GINI mutató értéke nagyon alacsony: 13 %.

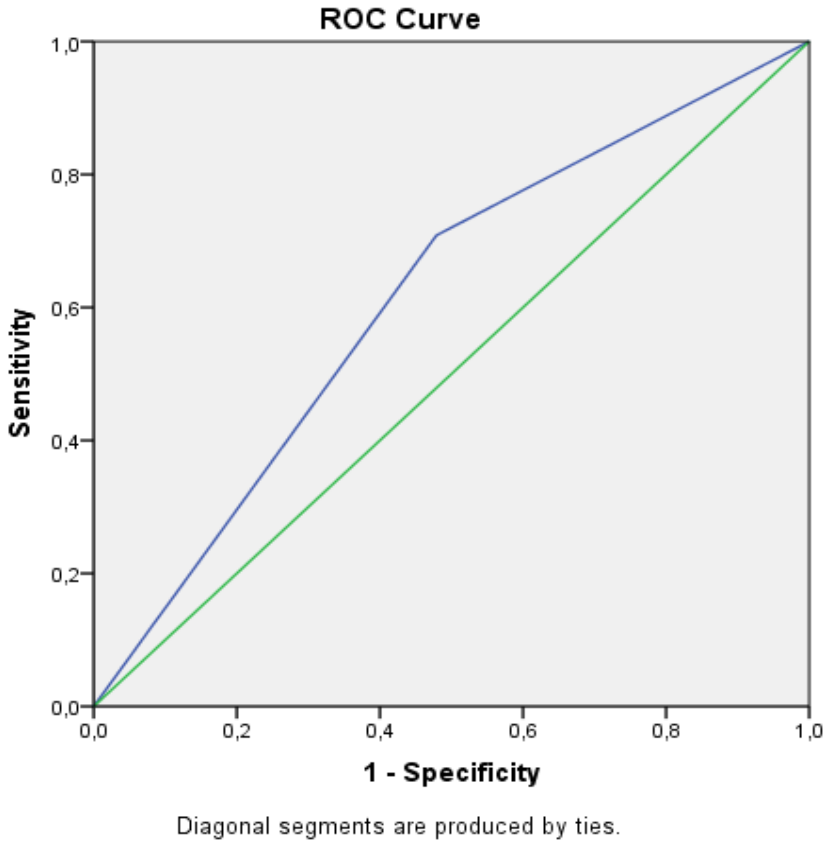
A medián alapú felosztás révén létrejövő bináris kategóriaváltozó paraméterbecslését és a ROC görbe alakulását tekinthetjük meg a következő táblázatban, illetve ábrán.

13. Táblázat. A „MEDIANKAT” változó paraméterbecslése

Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	mediankat(1)	,972	,270	12,929	1	,000	2,643
	Constant	-2,833	,225	159,204	1	,000	,059

16. ábra. A „MEDIANKAT” változó ROC görbéje



A GINI mutató értéke ebben az esetben még kisebb, értéke 23%.

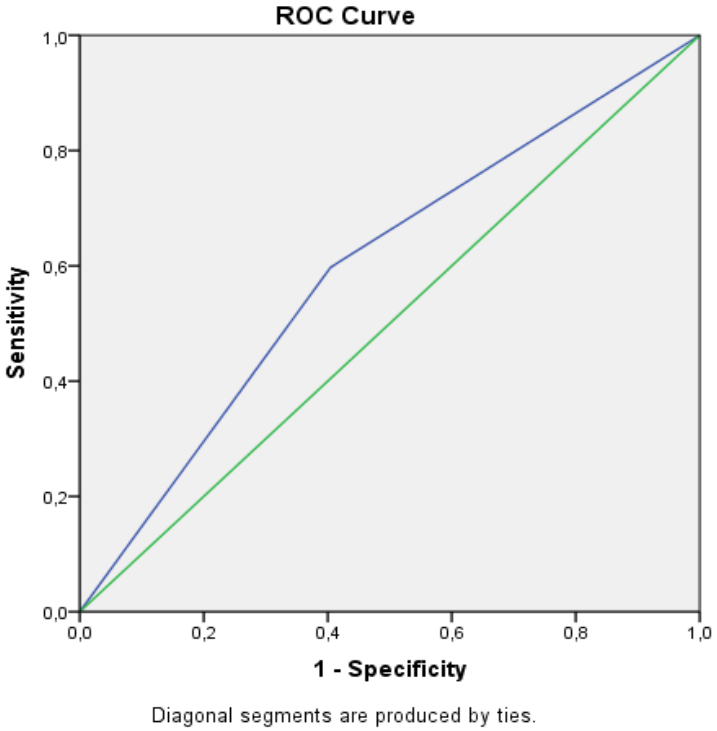
Végül az utolsó, az eloszlás átlaga által meghatározott „MEANKAT” változó esetében az eredmények:

14. Táblázat. A „MEANKAT” változó paraméterbecslése

Variables in the Equation

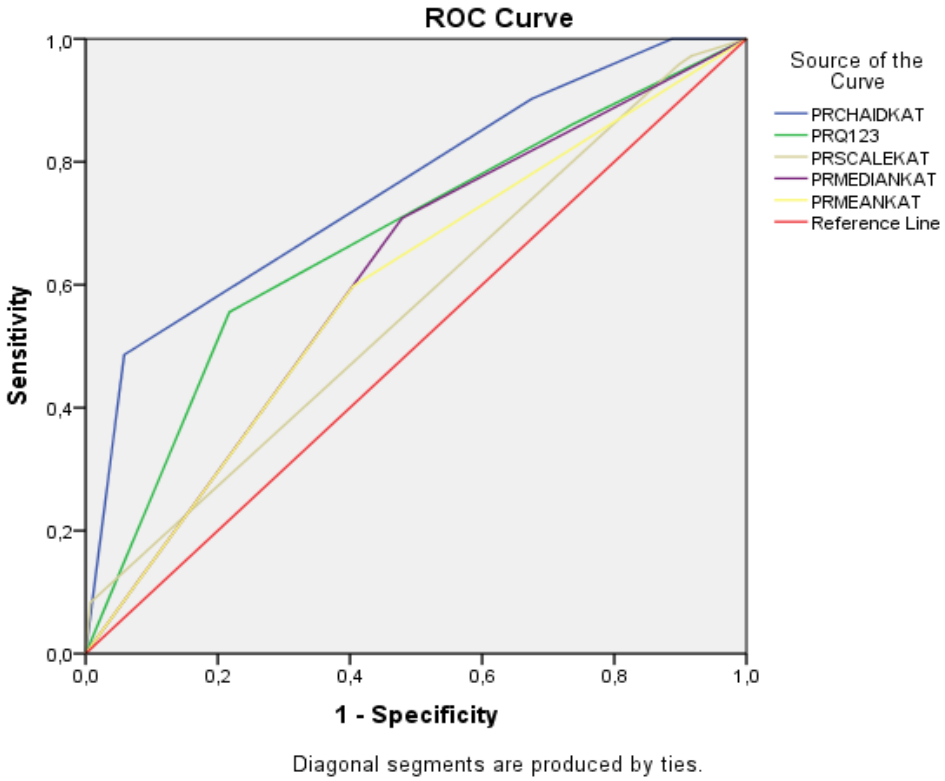
		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	meankat(1)	,781	,253	9,564	1	,002	2,184
	Constant	-2,644	,192	189,274	1	,000	,071

17. ábra. A „MEANKAT” változó ROC görbéje



Az illeszkedés nagyon gyenge a GINI mutató értéke 19,3%. A következő ábrán egyesítve láthatjuk azöt alternatív modell ROC görbéjét.

18. ábra. Egyesített ROC görbe.



Az ábrán látható, hogy a legnagyobb területet befoglaló görbe, ezáltal a legnagyobb prediktív erő az optimális kategorizálási eljáráshoz tartozó „CHAIDKAT" változóhoz tartozik. Az egyes modellekhez tartozó GINI értékeket az alábbi táblázatban foglaljuk össze.

15. Táblázat. GINI értékek összefoglaló táblázata

MUTATÓ	*CHAIDKAT*	*Q123KAT*	*MEDIANKAT*	*MEANKAT*	*SCALEKAT*
GINI	51,4%	34,7%	23%	19,3%	13%

VI/7.Folytonos változó hatványainak bevonása

A kategóriaegyesítési algoritmuseredményeként előálló új kategóriaváltozók kódolása az adatbázisban dummy-változók segítségével történik. Egy K kategóriával rendelkező változó esetében ez K-1 darab dummy-változó bevonását jelenti az adatbázisban[32]. A kategorizálással tehát az adatbázisban szereplőváltozók száma jelentősen megnőhet. Ennek következtében gyakran előfordul, hogy az egy változóra jutó esetszám (EPV) túl alacsony lesz. Ebben az esetben megoldást jelenthet, ha az újonnan létrehozott változók közül kiválasztjuk azokat amelyeknél az algoritmus kategória összevonást nem eredményezett, vagy csak egy-két kategória került összevonásra. Ezeknél a változóknál ugyanis a legmagasabb a létrejövő dummy-változók száma[33]. Ezeknél a változóknál a kategorizálás alternatívájaként, az eredeti folytonos változó megtartása mellett a változó négyzetének, esetleg harmadik hatványának egyidejű szerepeltetése segíthet a nem monoton jellegű kapcsolat kezelésében[34], az új változók számának alacsonyan tartása mellett.

[32] Azzal a megszokott feltételezéssel élve, hogy modellünk tengelymetszettel rendelkezik.
[33] Azoknál a változóknál, ahol kategória összevonást az algoritmus nem eredményezet K=10 esetén 9 db dummy-változó szükséges a változó kódolásához.
[34] Ez képletesen az egyszer „irányt váltó" kapcsolatoknak másodfokú, míg a kétszer irányt váltó kapcsolatoknak harmadfokú polinom segítségével történő közelítését jelenti.

VII. Valószínűségbecslés modellel

VII/1.A minta belső arányai

Az elkészült prediktív modellekkel kapcsolatban kézenfekvő elvárás, hogy segítségükkel egy adott esemény bekövetkezésének valószínűségét helyesen tudjuk becsülni. Mint arra korábban utaltunk a logisztikus regressziós modell segítségével csak akkor tudjuk torzítatlanul becsülni az előrejelezni kívánt esemény bekövetkezési valószínűségét, ha az „1" esemény bekövetkezésének mintabeli $\boldsymbol{y}$ aránya megegyezik a sokasági $\boldsymbol{P}$ (prior) aránnyal. A gyakorlatban előforduló elemzési szituációkban általában az egyezőség nem szokott fenn állni. Legtöbbször azért nem, mert az előrejelezni szándékozott kimenet jellemző ritkasága folytán a minta, már korábban tárgyalt módon, kiegyensúlyozatlanná válik, és ezért ritka egyedek további csatolására, vagy a gyakori kategória egyedeiből történő elhagyásra (case-control) van szükség. A minta eredeti belső aránya mindkét esetben meg fog változni, így kicsi az esély a sokasági eloszlás reprezentálására a célváltozó kategória arányainak szempontjából. A modell futtatását megelőzően az arányok tekintetében három eset lehetséges:

$$\mathbf{1}.\ \boldsymbol{y} < P \tag{7.1}$$

$$\mathbf{2}.\ \boldsymbol{y} = \boldsymbol{P}(\text{ez a legkevésbé valószínű eset}) \tag{7.2}$$

$$\mathbf{3}.\ \boldsymbol{y} > P \tag{7.3}$$

Amennyiben az egyezőség (2.eset) nem teljesül, a modell futtatását követően az alábbiakban ismertetésre kerülő úgynevezett torzításcsökkentő korrekció végrehajtására van szükség.

VII/2. Torzításcsökkentő prior korrekció

Egyszerű kivitelezhetősége folytán a leggyakrabban alkalmazott eljárás az úgynevezett torzításcsökkentő prior korrekció módszere (Prentice-Pyke, 1979). Az eljárás a

hagyományos maximum likelihood (ML) segítségével elkészült modell tengelymetszetét ($\boldsymbol{\beta_0}$) korrigálja a mintabeli és a sokasági arányok segítségével a következő módon:

$$\hat{\beta}_0 - \ln\left[\left(\frac{1-P}{P}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] = \hat{\beta}_0 - \left[\underbrace{\ln\left(\frac{\bar{y}}{1-\bar{y}}\right)}_{\hat{\beta}_{0_{ML}}} - \underbrace{\ln\left(\frac{P}{1-P}\right)}_{\beta_0}\right] \tag{7.4}$$

A formula segítségével a tengelymetszet korrigált konzisztens becslését kapjuk. A gyakorlati munka során ezek után csak annyit kell tennünk, hogy az elkészült modell paraméterei közül a $\boldsymbol{\beta_0}$tengelymetszeti paraméteren a korrekciót a fenti képlet segítségével elvégezzük. A torzítást csökkentő prior módszer segítségével a könnyen és gyorsan végrehajthatjuk a szükséges korrekciót az elkészült modellen. Az alkalmazás során azonban problémát jelent, hogy sok esetben semmilyen információval nem rendelkezünk a priori eloszlásra vonatkozóan. Ebben az esetben $\boldsymbol{P}$ értéke sem ismert. A következőkben röviden bemutatásra kerül egy olyan technika, amely a priori valószínűség ismeretének hiányában is alkalmas az esemény bekövetkezési valószínűségének megfelelő becslésére regressziós modell alkalmazása esetén.

VII/3.Gyakoriségi kalibráció

A priori eloszlás ismeretének hiányában a valószínűség becsléséhez a modell által predikált függvényértékek gyakorisági eloszlásából indulunk ki. A módszer alkalmazása során az egyes esetekre előrejelzett függvényértékeket először nagyság szerinti sorba rendezzük. Az így létrejövő eloszlást *K* számú osztóponttal egyforma számosságú függvényértéket tartalmazó osztályközökre bontjuk[35]. Ezek után minden egyes osztályközre kiszámoljuk az „1” esemény adott osztályközre jellemző relatív gyakoriságát, amit egyszerűen az adott osztályközben található „1” esemény előfordulásának és az összes osztályközbe eső eset hányadosaként értelmezünk. Az így keletkező *K*+1 darab relatív gyakoriság az adott osztályközökre jellemző tapasztalati valószínűségként értelmezhető. Első megközelítésben ezek után valószínűséget úgy becsülhetünk a modell segítségével, hogy először egy adott esetnél megfigyeljük, hogy a modell által predikált függvényérték melyik osztályközbe

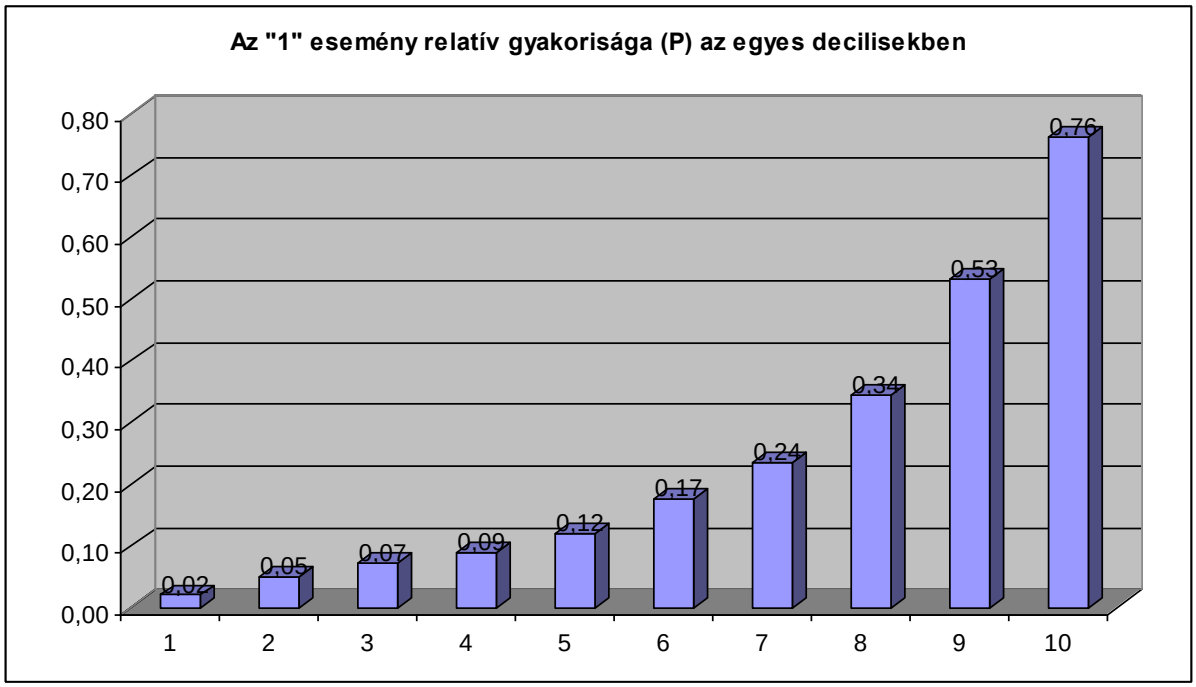
[35] *K*=9 esetén az eloszlás decilisekre bontásáról beszélünk

esik. Második lépésben az osztályközre jellemző relatív gyakoriságot hozzárendeljük az előrejelzett függvényértékhez. Továbbiakban ezt a relatív gyakoriságot tekintjük az adott függvényértékhez tartozó valószínűség becsléseként. A módszer hátránya, hogy adott osztályköz esetén ugyanazt a valószínűséget becsüljük az osztályköz legalacsonyabb és a legmagasabb függvényértéke esetén is. Ezzel összefüggésben az is problémát jelent, hogy összesen K+1 darab diszkrét valószínűségi becslésünk lesz, bármilyen nagy is legyen az előrejelezni kívánt sokaság elemszáma. Egy ilyen eloszlásra mutat logisztikus regressziós modell és K=9 esetén példát az alábbi táblázat és a hozzá tartozó ábra.

16. Táblázat. Relatívgyakoriságok az egyes decilisekben

Decilis	1	2	3	4	5	6	7	8	9	10
Relatív gyakoriság (P)	0,02	0,05	0,07	0,09	0,12	0,17	0,24	0,34	0,53	0,76

19. ábra. Relatív gyakoriságok az egyes decilisekben



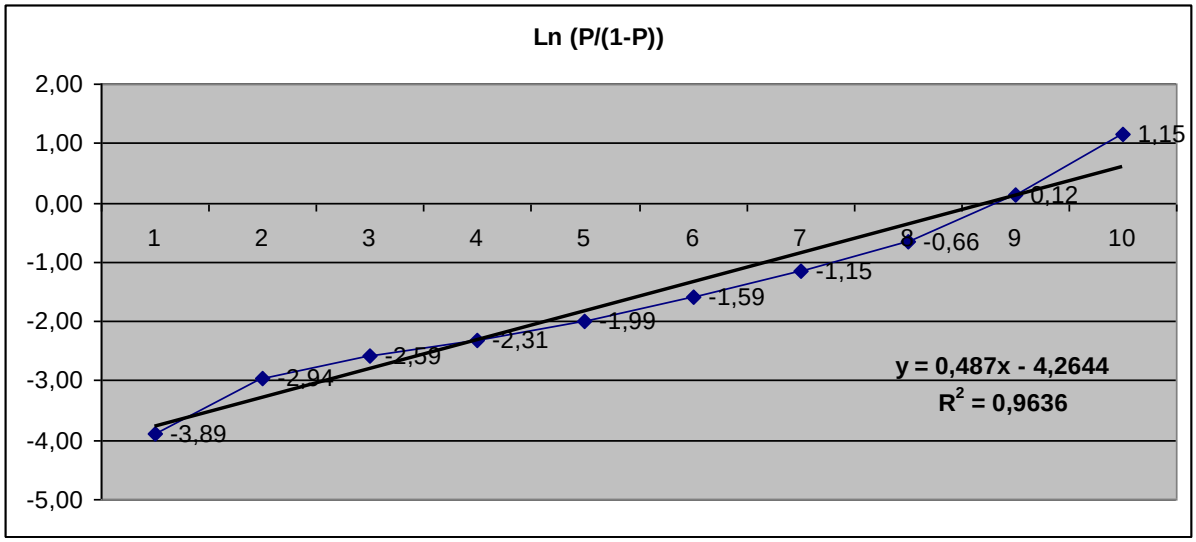
A gyakorlatban legtöbbször nem elégséges a valószínűségeknek ez a diszkrét értékkészleten megvalósuló becslése. Felmerül az igény, hogy eltérő modell által előrejelzett függvényértékhez különböző valószínűségi becslés tartozzon. A cél elérésére az osztályközökhöz tartozó relatív gyakoriságok, mint diszkrét valószínűségbecslésekre

történő függvényillesztés javasolható. Mint ismeretes, logit modell esetén az *odds (P/(P-1))* a magyarázó változók lineáris kombinációjaként áll elő. Ezt felhasználva illesszünk egyenest az relatív gyakoriságok ln (P/(1-P)) sorozatára. Jól illeszkedő logit model esetén magas R négyzetet várhatunk a lineáris illesztés során is amint azt az alábbi táblázat és a hozzátartozó ábra mutatja.

17. Táblázat. Ln(P/(1-P) értéke az egyes decilisekben

Decilis	1	2	3	4	5	6	7	8	9	10
Ln(P/(1-P)	-3,89	-2,94	-2,59	-2,31	-1,99	-1,586	-1,153	-0,663	0,12	1,15

20. ábra. Regressziós egyenes illesztése



A lineáris regressziós egyenlet (a példában: y = 0,487x - 4,2644) segítségével ezek után bármilyen logit függvényérték (x) esetén tudunk valószínűségi becslést adni (P). A módszer egyszerűsége és rugalmassága okán a gyakorlatban rendkívül elterjedt. A valószínűség becslés gyakoriségi kalibráción alapuló módszere például a pénzintézetek ügyfél nemfizetési valószínűség becslésénél a nemzetközi szabályozó hatóság (Basel Committee) részéről kötelező érvénnyel megkövetelt. Hátránya, hogy az illeszkedési hibák a két modell egymás utáni alkalmazása következtében „duplázódnak", ami különösen az eleve rosszul illeszkedő logit model esetén jelenthet problémát. Utaltunk már rá, hogy a módszer univerzális, bármilyen regressziós modell esetén kivitelezhető a fent leírtakhoz képest annyi különbséggel, hogy az utolsó lépésben, a lineáris regresszió illesztésénél az eredeti

prediktív regressziós modell által meghatározott *link-függvény* által meghatározott értékekre történik az illesztés.

VIII. Exploratív modellezés

A dolgozatban egészen idáig azt feltételeztük, hogy fő célkitűzésünk a lehető legjobban illeszkedő, legnagyobb előrejelzési erővel bíró prediktív modell megépítése. Az ilyen modellekre általában az a jellemző, hogy nagyszámú magyarázó változó szerepel bennük, amelyek tipikusan nem függetlenek egymástól. Számos előrejelzési helyzetben viszont a fő feladat az előrejelzendő esemény alakulását meghatározó fő tényezők meghatározása, értelmezése, egymáshoz viszonyított súlyuk megállapítása a modellezendő jelenség vonatkozásában. Az ilyen jellegű modellek, a csökkentett számú, és a magyarázhatóság miatt az egymás közötti függetlenség közelítése érdekében előszűrt magyarázó változók következtében, magyarázó erőben elmaradnak a maximálisan illeszkedő modellek mögött. A kétfajta modellezési cél közötti „trade-off"-ot már a kutatás megkezdése előtt célszerű figyelembe venni. A továbbiakban a logisztikus regresszió eredményeinek értelmezésén túl áttekintjük, hogy a modellezés folyamata mennyiben változik meg vagy egészül ki, ha célunk az exploratív modellezés.

VIII/1.Logisztikus regresszió paramétereinek értelmezése

Mint már korábban utaltunk rá a regressziók, így a logisztikus regresszió esetében, a magyarázó változók eredményváltozóra gyakorolt hatását akkor tudjuk egzaktul értelmezni, amennyiben a magyarázó változók egymástól függetleneknek tekinthetők. Ilyenkor nyílik lehetőség ugyanis az adott változóhoz tartozó béta paraméterek ceteris paribus értelmezésére. Tekintve, hogy az "odds" természetes alapú logaritmusa lineáris függvénye a magyarázó változóknak[36], ezért a paraméterek értelmezésénél a valószínűségekre gyakorolt közvetlen hatást nem, csak az odds-ra gyakorolt parciális (multiplikatív) hatást tudjuk interpretálni.

Ha X_j magyarázó változóhoz tartozó paraméter β_j, akkor X_j egységnyi növekedésének a hatására az "$odds_i$" várhatóan $e^{\beta j}$-szeresére növekszik ceteris paribus. Ebből az is látszik, hogy az így kiszámolt "odds" értéke a korábbi "odds"-nak is a függvénye.

[36] $odds_i = \exp(\beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \beta_3 * X_{i3} + \ldots + \beta_k * X_{ik})$

Az elmondottak szemléltetésére tegyük fel, hogy egy logisztikus regressziós modellben az X_1magyarázó változóhoz tartozó $\beta_1$paraméter becsült értéke $\beta_1=2.69$. Ekkor a X_1 változóban bekövetkező egységnyi növekmény hatására (feltételezve persze, hogy közben a többi magyarázó változó változatlan értékű) az odds várhatóan$e^{2.69}=14.73$-szeresére fog növekedni. Más szóval az X_1 változó egységnyi növekedése 14.73-szeresére növeli az "1" kategóriába tartozás valószínűségét a "0" kategóriába tartozás valószínűségéhez képest. Természetesen, ha a paraméter negatív értékű, ez azt jelenti, hogy a hozzá tartozó magyarázó változó egységnyi növekedése az "1" kategóriába tartozás valószínűségét csökkenti a "0" kategóriába tartozás valószínűségéhez képest. A paraméterek értelmezhetősége természetesen csak akkor lehetséges, ha igaz az, hogy a vizsgált változó megváltozása nem jár együtt más magyarázó változók megváltozásával.

A gyakorlati problémák többségében azonban a magyarázó változók nemegyszer szoros összefüggésrendszert, együttmozgást mutatnak. Ezt a jelenséget nevezzük a magyarázó változókra jellemző multikollinearitásnak. A multikollinearitás nemcsak azért jelent problémát, mert oksági kapcsolatok feltérképezését megnehezíti, esetleg lehetetlenné teszi, hanem mert a paraméterbecslések standard hibáját jelentősen növelheti a hagyományos lineáris és logisztikus regresszió alkalmazása során egyaránt[37]. Így, különösen kisebb minták esetén, fennáll a veszély, hogy releváns magyarázó változók nem bizonyulnak szignifikánsnak a szokásos szignifikancia szintek mellett, ezzel csökkentve a modell prediktív erejét. Ezért amikor a maximálisan előrejelző modell megalkotása a cél a mintaméret növelése (standard hibacsökkentő hatás) mellett a beléptetési és kiléptetési küszöb szignifikancia szintek emelésével (stepwise változószelekció alkalmazása esetében) lehet elérni, hogy egy adott változó a végleges modell része legyen. Természetesen ilyenkor a multikollinearitás következtében a modellezendő jelenség megértése, az oksági kapcsolatok feltárása pusztán a modellből általában nem lehetséges, modellünk fekete dobozként működik. Ilyenkor modellünk multikorrelált, a paraméterek értelmezése a fent leírt módon nem lehetséges, azonban a modell klasszifikációra továbbra is alkalmas. Mindebből az következik, hogy a magyarázó modell megépítésének feltétele az egymással nem korreláló magyarázó változók használata. A gyakorlatban sajnos ritkán fordul elő, hogy az induló magyarázó változó szett kölcsönösen független lenne egymástól. Ilyenkor

[37]Szélsőséges esetben, amennyiben a magyarázó változók közötti függvényszerű determinisztikus lineáris kapcsolat (egzakt multikollinearitás) jellemzi az adatszerkezetet, a maximum likelihood becslések logisztikus regresszió esetében nem lehetségesek, a program futása hibaüzenettel leáll.

egy lehetséges megoldás az eredeti változószettből egymással korrelálatlan változók azonosítása. Amennyiben ez nem lehetséges, célravezető lehet olyan egymással független úgynevezett látens változók (faktorok) lehetséges beazonosítása, melyek a kiinduló (indikátor) változók egy-egy csoportját reprezentálva alkalmasak a modellben történő szerepeltetésre. Ennek bemutatására az esettanulmányban térünk ki.

VIII/2. Változók standardizálása

Az egyes magyarázó változók hatásának, fent leírt módon történő, összehasonlítása problematikussá válhat, amennyiben a folytonos magyarázó változók eltérő mértékegységűek, azaz különböző skálán mérjük őket. Másképpen fogalmazva egy adott magyarázó változóhoz tartozó becsült paraméter és ezzel együtt az "odds"-ra gyakorolt hatása is megváltozik, amint a változót egy más mértékegységgel szerepeltetjük (pl. gramm helyett kilogramm).

A jelenség a dimenzió nélküli változóknál is megfigyelhető[38]. A magyarázó változók az előrejelzésben betöltött fontosságuk, súlyuk összehasonlítására rögtön lehetőség nyílik, ha a változókat "közös skálára" hozzuk, azaz standardizáljuk őket. A változók standardizálásának menete általában nagyon egyszerű. A leggyakrabban alkalmazott standardizálási eljárás, az úgynevezett *z*-score módszer esetén az adott *x* változó esetében a változó értékeiből a változó átlagának kivonása után a változó szórásával osztva kapjuk meg a standardizált értékeket az

$$z = \frac{x - \mu}{\sigma} \tag{8.1}$$

képletnek megfelelően. Ezek után a standardizált változókon épített modell paramétereinek abszolút érték sorrendjének megfelelően alakul az egyes releváns magyarázó változók előrejelzésben betöltött szerepük fontossága.

[38] A vállalati beszámolókból készített pénzügyi mutatók tipikusan ilyenek.

VIII/3.Esettanulmány

Csődelőrejelzés logisztikus regresszióval

Mintapéldánkban (Hámori, 2001) 757 létező vállalat segítségével készítünk csődelőrejelző modellt. A mintában szereplő vállaltok közel 10 százaléka fizetésképtelennek bizonyult. Az előrejelzéshez 21 darab, a vállalatok pénzügyi beszámolóiból készített mutatót használtunk fel. Az összehasonlíthatóság érdekében a mutatók változóit standardizáltuk. A végleges logisztikus regressziós modell jellemzőit[39] tartalmazza az alábbi táblázat:

18. Táblázat. Paraméterbecslés eredményei

Logistic Regression Table							
Odds	95% CI						
Predictor	béta	StDev	Z	P	Ratio	Lower	Upper
Constant	-5,5272	0,6093	-9,07	0,000			
köt/forr	1,2185	0,3113	3,91	0,000	3,38	1,84	6,23
aerd/árb	-1,2097	0,4975	-2,43	0,015	0,30	0,11	0,79
aerd/eszk	-2,2768	0,8712	-2,61	0,009	0,10	0,02	0,57
üerd/eszk	2,0042	0,7282	2,75	0,006	7,42	1,78	30,92
(ü+écs)/árbev	0,6908	0,2948	2,34	0,019	2,00	1,12	3,56
auerd/st	-1,0294	0,4023	-2,56	0,010	0,36	0,16	0,79
forge/rlk	-3,687	1,878	-1,96	0,050	0,03	0,00	0,99
köt/(ü+écs+pübev)	1,0750	0,2399	4,48	0,000	2,93	1,83	4,69
üerd/püráf	-0,9110	0,3159	-2,88	0,004	0,40	0,22	0,75
rlk/árb.	2,2918	0,5475	4,19	0,000	9,89	3,38	28,93
Log-Likelihood = -81,770 GINI=81,2%							
Test that all slopes are zero: G = 312,172; DF = 10; P-Value = 0,000							

A táblázat első és második oszlopa tartalmazza a prediktorokat a hozzájuk tartozó béta paraméterekkel együtt. A *p* értékeket tanulmányozva látszik, hogy mind a 10 magyarázó változó szignifikáns a szokásos szinteken. A modell globális illeszkedését tesztelő G-statisztika is a modell jó illeszkedését támasztja alá. Mivel a mutatók standardizáltak, ezért

[39]Az modellezés MINITAB programcsomag segítségével készült

a mutatókat paramétereik abszolút értékének nagysága szerint rendezve a következő sorrend alakul ki:

19. Táblázat. Mutatók sorrendje

1. Forgó eszközök / Rövid lejáratú kötelezettségek
2. Rövid lejáratú kötelezettségek / Nettó árbevétel
3. Adózás előtti eredmény / Eszközök
4. Üzemi eredmény / Eszközök
5. Kötelezettségek / Összes forrás
6. Adózás előtti eredmény / Nettó árbevétel
7. Kötelezettségek / (Üzemi eredmény+Értékcsökkenés+Pénzügyi bevételek)
8. Adózás utáni eredmény / Saját tőke
9. Üzemi eredmény / Pénzügyi ráfordítások
10. (Üzemi eredmény+Értékcsökkenés) / Nettó árbevétel

A mutatók fontossági sorrendje segít megválaszolni, hogy melyik mutatók és mennyire számítanak a fizetésképtelenség szempontjából. További fontos kérdés az, hogy hogyan függ az előrejelzésünk ezektől a mutatóktól. Erre a kérdésre viszont csak akkor tudnánk könnyen válaszolni, ha a felsorolt mutatók függetlenek lennének egymástól.

Korrelációvizsgálat

Vizsgáljuk meg tehát, hogy vajon a modell változói függetlenek-e egymástól. A magyarázó változók sztochasztikus kapcsolatának kimutatására a nemparaméteres Spearman-féle rangkorrelációtalkalmazzuk, tekintettel arra, hogy a rangkorreláció eredménye a mutatók eloszlásától független. A mutatók közötti rangkorrelációkat láthatjuk a táblázatban:

20. Táblázat. Korrelációs táblázat

		köt/ forr	rlk/ árb	aerd/ árb	aerd/ esz	üerd/ esz	(ü+écs)/ árb	auerd/st	forge/rl	köt/(ü+é)	üerd/pür
köt/forr		1,000	,379**	-,143**	-,011	,044	-,182**	,319**	,592**	,520**	,019
		,	,000	,000	,765	,222	,000	,000	,000	,000	,609
rlk/árb		,379**	1,000	,003	-,308**	-,283**	,135**	-,156**	-,491**	,442**	-,245**
		,000	,	,938	,000	,000	,000	,000	,000	,000	,000
aerd/árb		-,143**	,003	1,000	,800**	,712**	,799**	,673**	,148**	-,660**	,528**
		,000	,938	,	,000	,000	,000	,000	,000	,000	,000
aerd/esz		-,011	-,308**	,800**	1,000	,924**	,514**	,890**	,079**	-,716**	,697**
		,765	,000	,000	,	,000	,000	,000	,030	,000	,000
üerd/esz		,044	-,283**	,712**	,924**	1,000	,589**	,834**	,037	-,712**	,613**
		,222	,000	,000	,000	,	,000	,000	,310	,000	,000
(ü+écs)/ árb		-,182**	,135**	,799**	,514**	,589**	1,000	,404**,	,111**	-,630**	,301**
		,000	,000	,000	,000	,000	,	,000	,002	,000	,000
auerd/st		,319**	-,156**	,673**	,890**	,834**	,404**,	1,000	-,114**	-,456**	,654**
		,000	,000	,000	,000	,000	,000	,	,002	,000	,000
forge/rl		,592**	-,491**	,148**	,079**	,037	,111**	-,114**	1,000	-,326**	,048
		,000	,000	,000	,030	,310	,002	,002	,	,000	,186
köt/(ü+é)		,520**	,442**	-,660**	-,716**	-,712**	-,630**	-,456**	-,326**	1,000	-,444**
		,000	,000	,000	,000	,000	,000	,000	,000	,	,000
üerd/pür		,019	-,245**	,528**	,697**	,613**	,301**	,654**	,048	-,444**	1,000
		,609	,000	,000	,000	,000	,000	,000	,186	,000	,

***. Correlation is significant at the .01 level (2-tailed).*

A táblázat celláiban a felső szám jelzi a Spearman-féle rangkorreláció értékét, alatta, pedig a szignifikancia szint látható. A cella legalsó száma azt jelzi hány esetből számolta a program a korreláció értékét (esetünkben mindenhol 757 eset van, mivel nem volt adathiányos eset). A táblázatból kitűnik, hogy minden változó legalább 7 másik változóval korrelál. Ebből kifolyólag a logisztikus regresszió paramétereinek ceteris paribus értelmezése nem lehetséges.

A magyarázó változók számának csökkentése faktormodellel

A modell magyarázó változóinak vizsgálatakor, megállapítottuk, hogy azok egymással erősen korrelálnak, közöttük sztochasztikus kapcsolat van. Ennek alapján felvetődik a kérdés, hogy tudunk-e esetleg olyan, közgazdasági jelentéssel is bíró, korrelálatlan mesterséges (látens) változókat létrehozni, melyek az eredeti (indikátor) változók által hordozott információjelentős részét továbbra is megőrzik. Ebben az esetben ugyanis,

ezekkel a mesterséges változókkal újraépítve a modellt, lehetőség nyílna a paraméterek ceteris paribus értelmezésére.

A kívánt cél elérésére népszerű eszköz a faktoranalízis módszere. Egyszerű koncepcionális magyarázatként megfogalmazható, hogy a faktoranalízis alkalmazása során a cél a sokváltozós adatállomány jellemzése a változónál kisebb számú célszerűen választott mesterséges változókkal, az ún. faktorokkal. A faktoranalízis algoritmusa ehhez egy olyan optimális súlyrendszert keres, melyek segítségével az eredeti változók a kívánt számú látens faktor lineáris kombinációjaként legjobban reprodukálhatók. A faktoranalízis alkalmazása során kirajzolódnak azok, az eredeti változókból meghatározott változócsoportok, melyek egymással erősen korrelálnak úgy, hogy az egyes csoportok a hozzájuk tartozó faktorokkal reprezentálódnak és a származtatott faktorok egymás között tipikusan lineárisan függetlennek tekinthetők. A módszer kimerítő leírását tartalmazza Hajdu (2003) munkája.

Konkrét elemzésünk során, az analízis részleteit mellőzve az alábbiakban csak a legfontosabb eredmények kerülnek bemutatásra. A faktoranalízis során a 10 standardizált alapváltozóból 4 faktor került kialakításra, a faktorok rotálása[40], pedig varimax módszerrel történt:

[40] Az eredeti faktorok, mint főkomponensek az információ csökkenő hányadát reprezentálják (lásd. Variance sor értékei). A rotálás célja az információ egyenletesebb elosztása az egyes faktorokon.

21. Táblázat. Rotálatlan faktorsúly mátrix

Principal Component Factor Analysis of the Correlation Matrix					
Unrotated Factor Loadings and Communalities					
Variable	Factor1	Factor2	Factor3	Factor4	Communality
köt/forr	0,111	-0,355	0,856	0,221	0,919
aerd/árbe	-0,600	0,630	0,325	0,013	0,863
aerd/eszk	-0,889	-0,311	-0,154	-0,005	0,911
üerd/eszk	-0,864	-0,287	-0,126	-0,022	0,846
(ü+écs)/árbe	-0,582	0,582	0,310	0,001	0,774
auerd/st	-0,727	-0,492	0,285	0,128	0,867
forge/rlk	0,011	0,165	-0,350	0,917	0,991
köt/(ü+écs+pübe	0,726	-0,108	0,433	0,182	0,759
üerd/püráf	-0,631	-0,341	0,019	0,109	0,527
rlk/árb.	0,424	-0,622	-0,136	-0,048	0,588
Variance	3,8811	1,8250	1,3835	0,9543	8,0440
% Var	0,388	0,183	0,138	0,095	0,804

22. Táblázat. Rotált faktorsúly mátrix

Rotated Factor Loadings and Communalities					
Varimax Rotation					
Variable	Factor1	Factor2	Factor3	Factor4	Communality
köt/forr	0,075	0,020	-0,951	0,090	0,919
aerd/árbe	0,165	-0,914	0,005	0,021	0,863
aerd/eszk	0,922	-0,146	0,199	0,017	0,911
üerd/eszk	0,886	-0,163	0,182	0,037	0,846
(ü+écs)/árbev	0,175	-0,861	0,005	0,034	0,774
auerd/st	0,872	-0,085	-0,313	0,042	0,867
forge/rlk	-0,013	-0,021	0,067	-0,993	0,991
köt/(ü+écs+pübev	-0,562	0,288	-0,599	-0,045	0,759
üerd/püráf	0,721	-0,059	-0,045	-0,042	0,527
rlk/árb.	-0,030	0,752	-0,127	0,071	0,588
Variance	3,2932	2,2855	1,4576	1,0077	8,0440
% Var	0,329	0,229	0,146	0,101	0,804

Az első táblázat a rotálatlan faktorsúly-mátrix. Emlékeztetőül a jobb szélső oszlop tartalmazza a kommunalitásokat, melyek azt mutatják, hogy az adott alapváltozó varianciájának összesen hány százalékát magyarázzák meg a faktorok. Látható, hogy a 4

faktor együttesen a 10 alapváltozó varianciájának több mint 80%-át magyarázza meg.A következő táblázat a rotált faktorsúly-mátrix. A rotálás célja az információ egyenletesebb elosztása a 4 faktoron. Megfigyelhető, hogy jól használható faktor-struktúra alakult ki (sárgával jelölve), azaz minden változó csak egy faktorban szerepel nagy súllyal (erősen korrelál vele), ez alól kivételt csak a "Kötelezettségek / (üzemi eredmény+écs+pénzügyi bevétel)" mutató, amely mind az első (-0.562), mind a harmadik faktorban (-0.599) viszonylag nagy súllyal szerepel. Továbbá mindegyik faktorhoz tartozik olyan változó, melynek sorában éppen ehhez a faktorhoz tartozik a legnagyobb a súly (a táblázatban sárgával jelölve).Így az eredeti 10 változó információ tartalmának 80%-át 4 egymással korrelálatlan faktorba sikerült tömöríteni az alábbi táblázatokban látható módon:

23. Táblázat. Első faktor mutatói

Faktor1:
Adózás előtti eredmény / Eszközök
Üzemi eredmény / Eszközök
Adózás utáni eredmény / Saját tőke
Üzemi erdmény / Pénzügyi ráfordítások

24. Táblázat. Második faktor mutatói

Faktor2:
Adózás előtti eredmény / Nettó árbevétel
(Üzemi eredmény+érték csökkenés) / Nettó árbevétel
Rövid lejáratú kötelezettségek / Nettó árbevétel

25. Táblázat. Harmadik faktor mutatói

Faktor3:
Kötelezettségek / Források
Kötelezettségek / (Üzemi eredmény+Értékcsökkenés+Pénzügyi bevételek)

26. Táblázat. Negyedik faktor mutatói

Faktor4:
Forgó eszközök / Rövid lejáratú kötelezettségek

A faktoroknak szemléletes jelentést is tulajdoníthatunk. Az első faktor csoportjába csak olyan mutató tartozik, amelynek a vállalt eredményességét hivatott kifejezni. Ennek alapján

ezt a faktort nevezhetjük eredményességfaktornak. A második, illetve a harmadik faktornál is hasonlóképpen járhatunk el: A második faktornál minden mutató nevezőjében az árbevétel, a harmadik faktornál a számlálókban csak a kötelezettségek szerepelnek. Ennek megfelelően legyen ez a két faktor az árbevétel-, valamint az eladósodottságfaktor. Az utolsó faktorhoz kizárólag csak a likviditási mutató tartozik, ezért ezt a faktort likviditásfaktornak nevezzük el.

Logisztikus regresszió faktorokkal

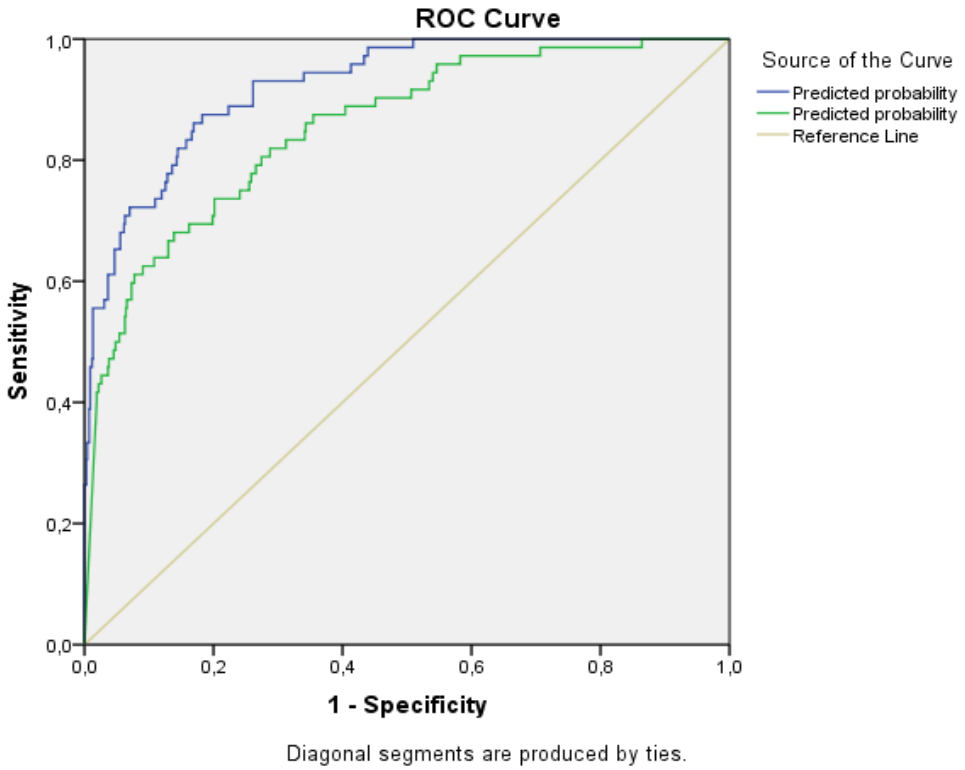
A faktoranalízis eredményeképpen minden mintabeli vállalat rendelkezik a négy faktorban felvett értékekkel.
Kézenfekvő lépés a logisztikus regressziós modell 4 független faktor segítségével történő újraépítése. Az eredményeket az alábbiakban láthatjuk:

27. Táblázat. Paraméterbecslés faktorokon

Logistic Regression Table							
Odds	95% CI						
Predictor	Coef	StDev	Z	P	Ratio	Lower	Upper
Constant	-5,7014	0,5571	-10,23	0,000			
Faktor1	-3,0113	0,4193	-7,18	0,000	0,05	0,02	0,11
Faktor2	1,1065	0,4270	2,59	0,010	3,02	1,31	6,98
Faktor3	-2,0891	0,2692	-7,76	0,000	0,12	0,07	0,21
Faktor4	5,574	1,175	4,74	0,000	263,46	26,32	2636,85
Log-Likelihood = -97,081 GINI=74,8%							
Test that all slopes are zero: G = 276,831; DF = 4; P-Value = 0,000							

A faktorokhoz tartozó paraméterek együttesen és külön is legalább 1%-os szinten szignifikánsak, a modell illeszkedése gyengébb, mint a tízváltozós esetben, de még mindig nagyon jó.A modell illeszkedését leíró GINI mutató faktorokon épített modell esetében 72% szemben az eredeti tízváltozós modell 83% értékével. A modellek illeszkedése közötti különbséget az alábbi ROC diagramon követhetjük nyomon:

21. ábra. ROC görbe két modellre



Az illeszkedés gyengülése annak a következménye, hogy a faktorok szerepeltetésével elvesztettük a tíz alapváltozó által hordozott információ közel 20%-át.A módszer nagy előnye ott érhető tetten, hogy ugyan a modell illeszkedése gyengébb, de most már lehetőség nyílik a magyarázó változóknak (faktorok) a csőd valószínűségére gyakorolt hatásának (ceteris paribus) elemzésére, valamint az egyes faktoroknak, a csőd kialakulásában betöltött súlyának összehasonlítására. A faktorok ugyanis korrelálatlanok, minek következtében helyénvaló az a feltételezés, hogy az egyik faktor ceteris paribus megváltozik. A faktorokat, pedig azáltal, hogy standardizáltak ("közös mértékegységen vannak"), a csődre gyakorolt hatásuk nagyságát paramétereik abszolút értéke alapján becsülhetjük. Ennek alapján a faktorok a csőd kialakulásában betöltött szerepük alapján sorrendbe állíthatók:

1. $Faktor4$ - $(\beta_4 = 5{,}574)$ (8.2)
2. $Faktor1$ - $(\beta_1 = -3{,}011)$ (8.3)
3. $Faktor3$ - $(\beta_3 = -2{,}089)$ (8.4)
4. $Faktor2$ - $(\beta_2 = 1{,}11)$ (8.5)

A mintában a csőd kialakulásában a legnagyobb szerepe tehát a likviditásfaktornak van. A paraméter pozitív értéke azt jelzi, hogy a faktor növekedése esetén a csőd valószínűsége növekszik a "nemcsőd" bekövetkezési valószínűségéhez képest. A faktor erős negatív sztochasztikus kapcsolatban van a likviditási mutatóval (korreláció= - 0,993), tehát a faktor növekedése mögött általában a likviditási mutató csökkenése áll.

A második helyen álló eredményességfaktor paraméterének negatív értéke miatt, a faktor növekedésére a csőd valószínűsége csökken a "nemcsőd" bekövetkezési valószínűségéhez képest. Tekintve, hogy a faktor az eredmény típusú mutatókat tartalmazza (méghozzá mindegyikkel erősen pozitívan korrelálva), ezért növekedése az eredményesség növekedését jelzi.

A sorrendben a harmadik az eladósodottság faktor. A hozzá tartozó paraméter negatív értékének ugyanaz a jelentése, mint az eredményességfaktornál: növekedése esetén az "odds" csökken. A faktor és a hozzárendelt mutatók között erős negatív kapcsolat van, azaz a faktor értékének növekedése az eladósodottsági mutatók csökkenésével jár együtt.

Az utolsó, árbevételfaktor, a két az eredményességet az árbevétel arányában kifejező mutatóval negatív, a rövid lejáratú kötelezettség/árbevétel mutatóval erős pozitív kapcsolatban van. A hozzá tartozó paraméter pozitív, ami azt jelenti, hogy növekedésével (az eredmény/árbevétel mutatók csökkenésével, rövid lejáratú kötelezettség/árbevétel mutató növekedésével) az "odds" nő.

A faktorok értékében bekövetkező változások "odds"-ra gyakorolt hatását a paraméterek alapján becsülhetjük. Az árbevételfaktor egységnyi növekedése például a csőd bekövetkezésének valószínűségét a "nemcsőd" bekövetkezési valószínűségéhez képest $e^{1,11}$-szeresére növeli.

Értelmezés

A faktorokon felépült modell további előnye, hogy lehetővé teszi egy adott eset (vállalat) tömör jellemzését. Példaképpen tekintsünk két vállalatot, A-t és B-t. A két vállalat kiinduló 10 változós logisztikus regressziós modelljének egyes magyarázó változóinak eredeti (nem standardizált) felvett értékeit, valamint a modell által előrejelzett függvényértéket mutatja be az alábbi táblázat.

28. Táblázat. Vállalatok mutatói

Mutatók	**"A" vállalat**	**"B" vállalat**
Kötelezettségek / összes forrás	0,72	0,50
Üzemi eredmény / összes eszköz	0.78	0,20
Adózás előtti eredmény / összes eszköz	0,017	0,22
Adózás előtti eredmény / nettó árbevétel	0,11	0,26
Forgóeszközök/rövidlejáratú kötelezettségek	0.78	1.36
Üzemieredmény/pénzügyi ráfordítások	1,42	3.23
Kötelezettségek/(üzemi eredmény + értékcsökkenési leírás + pénzügyi bevételek)	5.56	1.76
Rövid lejáratú kötelezettségek / nettó árbevétel	0.41	0.44
(Üzemi eredmény + értékcsökkenési leírás)/nettó árbevétel	0.09	0.23
Adózás utáni eredmény / saját tőke (ROE)	0.06	0.46
Előrejelzett függvényérték	***0.66***	***0.005***

Az előrejelzés alapján az A vállalat csődközeli helyzetben van (0.66), míg a B vállalatot a fizetésképtelenség nem fenyegeti (0.005). A két vállalat mutatóit külön-külön vizsgálva feltűnhet, hogy az egyes mutatókban nem mutatkozik olyan mértékű különbség a két vállalat között, ami alapján ilyen nagy előrejelzési különbséget várnánk. Pusztán az alapmodell segítségével nem igen tudunk magyarázatot adni jelentős előrejelzési különbségre. A két vállalatról egészen más képet kapunk, ha megvizsgáljuk, hogy milyen értékek jellemzik őket a faktorok szempontjából. Az alábbi táblázat tartalmazza a két vállalatnak a négy faktorban felvett értékét:

29. Táblázat. Faktorok értékei

Faktorok	**A vállalat**	**B vállalat**
Faktor1 (eredményesség)	-0.839	0.14
Faktor2 (árbevétel)	-0.136	-0.38
Faktor3 (eladósodottság)	-0.83	-0.03
Faktor4 (likviditás)	0.314	0.317

A faktorok mindegyikének, a teljes mintán értelmezett eloszlásában, nulla átlaga és egységnyi szórása van. Ennek következtében, egy adott esetben a faktor értékének negatív mivolta azt jelenti, hogy az adott eset átlag alatti, ellenkező esetben, pedig, hogy átlag

feletti. Látható, hogy az eredményesség faktor tekintetében az A vállalat jóval a minta átlaga alatt van, míg a B vállalat átlag feletti.

Az árbevétel faktor mindkét esetben átlag alatti értékű, és mint ilyen, ez mindkét vállalatnál kedvező, hiszen tudjuk, hogy ennek a faktornak a növekedése a fizetésképtelenség bekövetkezésének valószínűségét növeli a "nemcsőd" bekövetkezési valószínűségéhez képest (β_2 = 1,11). Az eladósodottság tekintetében, megint nagy különbség mutatkozik a két vállalat között. Az eladósodottság tekintetében az A vállalat jóval átlag alatti, míg a B vállalat átlagos értékkel rendelkezik. Emlékezzünk vissza, hogy itt a faktor magasabb értékei jelentették a vállalat kisebb mértékű eladósodottságát, és ebből kifolyólag az alacsonyabb "odds"-ot (β_3 = -2,089).

A negyedik faktor (likviditás) tekintetében nem mutatkozik lényeges különbség a két vállalat között. Pozitív értéke azt jelenti, hogy a likviditási mutató tekintetében a minta átlagához képest (mintaátlag=2.62) a két vállalatátlag alatti.

Összefoglalva megállapítható, hogy a két vállalat közötti különbség az eredményesség és az eladósodottság tekintetében érhető tetten. "A" vállalat eredményessége lényegesen rosszabb "B" vállalat eredményességénél, valamint sokkal jobban el van adósodva. Ez a két faktor magyarázza a magas függvényértéket (0.66).A fentiekből az is látszik, hogy a faktorok segítségével nemcsak két vállalat összehasonlítására nyílik lehetőség, hanem a vállalatokat össze tudjuk vetni az "átlagos" vállalattal (mind a négy faktorban zérusértékű) is. Természetesen az "átlagos" vállalat jellemzői attól függnek, hogy milyen annak a mintának az összetétele, melyen a modellt felépítettük. Esetünkben ugyanis a 757 elemű mintába 685 db olyan cég került be, melyek jelenleg nincsenek csőd közeli helyzetben, és ezért a teljes mintán értelmezett átlagok nyilván sokkal közelebb vannak ezen cégek átlagaihoz, mint a csődös cégek átlagaihoz. Ebből az is következik, hogy az "átlagos" vállalatnak igen kicsi a csődbe jutási valószínűsége.[41]

Konkluzió

Az esettanulmány segítségével bemutatásra került, hogy milyen lépések lehetnek szükségesek ahhoz, hogy modellünk segítségével nemcsak a csődesemény megfelelő

[41]Az "átlagos" vállalat bedőlési valószínűségének kiszámításakor, tekintve, hogy a faktorok mind zérus értékűek, a függvény exponensében csak a konstans tag fog szerepelni: $P = \exp(\beta_0) / (1+\exp(\beta_0)) = \exp(-5,7) / (1+\exp(-5,7)) = 0.0033$

előrejelzésére nyíljon lehetőség, hanem a fizetésképtelenség bekövetkezéséhez vezető tényezőket azonosíthassuk, kapcsolatrendszerüket feltárhassuk. A magyarázó (faktorokon épített) modell előrejelző ereje gyengébb, ami elsősorban a kisebb számú prediktor modellben történő szerepeltetésének tudható be. Cserébe a standardizált faktorok paramétereinek abszolút értékeinek sorrendje tükrözi a faktorok csődesemény bekövetkezésében betöltött súlyát, és a független faktorok segítségével a paraméterek odds-ra gyakorolt multiplikatív hatása is értelmezhetővé válik.

A gyakorlatban nem mindig nyílik lehetőség az esettanulmányban bemutatott faktormodell kialakítására. Ilyen esetben célszerű az előzetes korreláció-vizsgálat, illetve kategória változók esetén függetlenségvizsgálat[42]alkalmazásával az egymással legkevésbé összefüggő magyarázó változókat előszűrni, és a végleges modell építését ezeken a változókon elkezdeni. Az előszűrés kritériumát ebben az esetben is a célváltozóval történő kapcsolat szorossága adja.

[42]Vegyes (folytonos-kategóriás mérési szint esetén) kapcsolat esetén asszociáció vizsgálat alkalmazásával

Új, illetve újszerű tudományos eredmények

A regressziós modellezés gyakorlatához kapcsolható technikai részletkérdések külön-külön nagyon jól feltérképezett területnek tekinthetők, az egyes problémákra fokuszálva nagyon sok mérvadó tanulmány született az elmúlt évtizedekben. A szerzők általában a kérdések mögött meghúzódó matematikai háttér igényes bemutatásával és igénybevételével izoláltan, a modellkészítés folyamatából kiragadva, elemezik az egyes részletterületeket, problémákat. Az alkalmazott statisztikai modellezés, a szélesebb értelemben vett adatbányászati megoldások rohamos és iparszerű elterjedésével azonban megjelent egy új igény az egyes, a modellfejlesztés során felmerülő kérdéseket a modellezés folyamatában vizsgáló és elemző megközelítés iránt. A dolgozat a predikciós célú bináris klasszifikációk részterületén, a logisztikus regressziós modellt a vizsgálódás középpontjába helyezve, tesz kísérletet ennek a statisztika területén újszerű szemléletmódnak az érvényesítésére. A tanulmány segítségével átfogó képet kapunk a teljes modellezési folyamatról, az egyes lépéseknél jelentkező tipikus döntési helyzetekről, és az azokra adható válaszokról, legjobb gyakorlatokról. A hivatkozások segítségével az egyes részletkérdések iránt mélyebb ismeretre vágyók is útmutatást kapnak. A dolgozat így mind a gyakorló, mind az elméleti szakemberek érdeklődésére is igényt tarthat.

A disszertáció elején megfogalmazásra kerülő hipotézisek vizsgálata további értékét adhatja a dolgozatnak. Sikerült igazolni, hogy a szélsőséges értékkel rendelkező megfigyelések megfelelő kezelésével a regresszió illeszkedése javítható ***(hipotézis 1)***. Ugyanígy kimutatásra került, hogy a folytonos változók esetében a változó kategorizálása még abban az esetben is javíthatja a modell illeszkedését, ha egyébként a célváltozó és a folytonos prediktor közötti kapcsolat jellege monoton természetű ***(hipotézis 2)***. Ugyanitt bemutattunk egy az illeszkedést befolyásoló, és a disszertációban bemutatott vizsgálat által az adatmintán igazoltan optimális kategóriahatárok meghatározására alkalmas módszert ***(hipotézis 3).***

Az esettanulmány segítségével megmutattuk, hogy a maximális magyarázó erővel bíró predikciós modell esetében, a nagyszámú prediktor között általában meglévő jelentős multikollinearitás következtében, a regresszió paramétereinek ceteris paribus értelmezésére nincs lehetőség, az így elkészült modell általában fekete dobozként működik. Fordítva, a

független változószetten (faktorokon) felépülő modell illeszkedése gyengébb, viszont lehetőség nyílik a magyarázó változók, így az esemény bekövetkezését befolyásoló tényezők értelmezésére, kapcsolatrendszerük értékelésére.

Köszönetnyilvánítás

Az értekezés elkészülésében nagy szerepet játszott témavezetőm, Száz János Professzor Úr, aki értékes megjegyzésivel és bátorító gondolataival mindig átsegített a nehéz időszakokon, további lendületet adva az alkotó munkának. Köszönöm Neki. Hasznos útmutatásokat köszönhetek Hunyadi László professzor Úrnak is, aki a tőle megszokott alapossággal véleményezte az elkészült fejezeteket. Hálával gondolok a szerkesztési munkában nyújtott segítségéért Mile Boglárkára is. És végezetül, de nem utolsóként a sorban meg kell köszönnöm Dr. Oravecz Beatrixnak, feleségemnek a sok értékes szakmai segítséget, valamint azt, hogy a disszertáció írása során a nyugodt, alkotó légkört otthonunkban megteremtette.

Irodalomjegyzék

Paul Allison (2013): „ What is the best R-squared for logistic regression?”, *StatisticalHorizons, http://www.statisticalhorizons.com/r2logistic*

Paul Allison (2013): „ Missing data”, *Statistical Horizons,* http://www.statisticalhorizons.com/wp-content/uploads/2012/01/Milsap-Allison.pdf

Andridge, R.R-Little, R.J.A. (2010): *A Review of Hot Deck Imputation for Survey Non-response,International statistical review,* 78 40-64

V.Barnett – T.Lewis (1994): „Outliers is statistical data”, *John Wiley & Sons,* New York.

J.A.Bilmes (1998): „A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models”, *International Computer Sciences Institute,* http://melodi.ee.washington.edu/people/bilmes/mypapers/em.pdf

Paul R. Brown (1983): „Independent auditor judgment int he evaluation of internal audit functions”, *Journal of accounting research* 21 444-455

Alfonso Croeze- Lindsey Pittman-Winnie Reynolds (2012): „Solving nonlinear least squares problems with the Gauss-Newton and Levenberg Marquardt Methods”, Lousiana State University 1-16

G. Chen-T.Astebro (2001):„The Economic Value of Reject Inference in Credit Scoring,” *Presented at the conference of credit scoring and credit control*, Credit Research Centre, University of Edinburgh

A.P. Dempster, N.M. Laird és D.B. Rubin (1977):„Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm,” *Journal of the Royal Statistical Society B*, 39. évf. 1-38

S.G. Donald (1995):„Two-step Estimation of Heteroskedastic Sample Selection Models,” *Journal of Econometrics*, 65. évf. 347-380

J.W. Graham és S.I. Donaldson (1993):„Evaluating Interventions with Differential Attrition: the Importance of Nonresponse Mechanisms and Use of Followup Data,” *Journal of Applied Psychology*, 78. évf. 119-128

J.W: Graham (1996): „Missing Data:Analysis and design”,*Statistics for social and behavioral sciences* 178-179

R. Glynn, N. M. Laird és D.B. Rubin (1986): „Selection modeling versus mixture modeling with nonignorable nonresponse”, *In H. Wainer [ed.] Drawing Inferences from Self-Selected Samples*, 119-146. New York: Springer-Verlag.

D.J. Hand (2001):„Measuring Diagnostic Accuracy of Statistical Prediction Rules," *Statistica Neerlandica*, 53. évf. 3-16

Hámori Gábor (2001): „Fizetésképtelenség előrejelzése logit-modellel". *Bankszemle. / 1-2*.65-87

Hámori Gábor (2001): „Chaid-alapú döntési fák jellemzői". *Statisztikai Szemle* 79. évfolyam / 8 703-710

Hajdu Ottó (2004): „A csődesemény logit-regressziójának kismintás problémái." Statisztikai Szemle 4, 390-422

Hajdu Ottó (2003): „Többváltozós statisztikai számítások." Központi Statisztikai Hivatal, Budapest.

Hajdu Ottó és Virág Miklós (1996): „Pénzügyi mutatószámokon alapuló csődmodell-számítások", *Bankszemle,* 1996/1-2

Y.Haitovsky (1968): „Missind data is regression analysis", *Journal of the royal statistical society, series B,* 30 67-82

D.J. Hand- W.E. Henley (1993/4): „Can Reject Inference Ever Work?," *IMA Journal of Mathematics Applied in Business & Industry*, 5/4. 45-55.

D.M:Hawkins(1980):„Identification of outliers", *Chapman and Hall,* London

J. Heckman (1979):„Sample Selection Bias as a Specification Error," *Econometrica*, 47. évf. 153-161

D. Hedeker és R. D. Gibbons (1997): „Application of random-effects pattern-mixture models for missing data in longitudinal studies",*Psychological Methods,*2[1], 64-78

D. Holt, T.M.F. Smith és P.D. Winter (1980):„Regression Analysis of Data From Complex Surveys," *Journal of the Royal Statistical Society*, 143. évf. A.

M.Y.Hu-G.Zhang-D.Indro (1999): „*Artificial neural networks in bankruptcy prediction:General framework and cross-validation analysis*", European Journal of Operational Research, 114 évf. 16-32

Paul J. Hoffman-Paul Slovic- Leonard G.Rorer (1968): „An analysis-of-variance model for the assessment of configural cue utilization in clinical judgement", *Psychological Bulletin* 69 338-339.

Hosmer, D.W.- S. Lemeshow (2000):„Applied logistic regression", *John Wiley and Sons*, 373

Hunyadi László -Vita László (2002): „Statisztika közgazdászoknak", *Központi Statisztikai Hivatal,* Budapest

Daniel Kahneman (2013): „Gyors és lassú gondolkodás”, *HVG Kiadói Rt.*257-270

J.O. Kim és J. Curry (1977):„The treatment of missing data in multivariate analysis”, *Sociol. Meth. Res.,* 6.évf. 215-240

Kiss Ferenc (2003):„A credit scoring fejlődése és alkalmazása”, *Ph.D. értekezés,* BME

P.W. Lavori, R. Dawson és D. Shera (1995): „ A multiple imputation strategy for clinical trials with truncation of patient data”,*Statistics in Medicine*, 14 évf., 1913-1925

S. Y. Lee, Y.M. Chiu (1990): “Analysis of Multivariate Polychoric Correlation Models with Incomplete Data”, *British Journal of Mathematical and Statistical Psychology*, vol. 43, 145-154

R.J.A. Little (1993):„Pattern-mixture Models for Multivariate Incomplete Data,” *Journal of the American Statistical Association*, 88. évf. 125-134

R.J.A. Little- D.B. Rubin (2002):„Statistical Analysis with Missing Data”, *John Wiley & Sons*, 2nd edition, New York

R.J.A. Little- D.B. Rubin (1987):„Statistical Analysis with Missing Data”, *John Wiley & Sons*, New York.

Máder Miklós Péter (2005):„Imputálási eljárások hatékonysága”, *Statisztikai Szemle*, 83.évf. 7.szám, 628-644

M.L. Marais, J.M. Patell és M.A. Walfson (1984):„The Experimental Design of Classification Models: An Application of Recursive Partitioning and Bootsrapping to Commercial Bank Loan Classifications”, *Journal of Accounting Research*, 1984 Vol.22, 87-115

Yiftach Nagar- Thomas Malone (2011): „Combining human and machine intelligence for making predictions”, *MIT center for collective intelligence working paper No.2011-02* 1-6

Oravecz Beatrix (2008): „Hiányzó adatok és kezelésük a statisztikai elemzésekben”, *Staisztikai szemle*, 86.évfolyam 4.szám 365-383

K. Tiwari-K. Mehta- N.Jain (2007): „Selecting the appropriate outlier treatment for common industry applications”, *Nesug 2007 statistics and data analysis*, 1-5

P.L.Roth (1994):„Missing data: a conceptual review for applied psychologists”, *Personnel psychology* 47, 1-24

P. J. van der Loo (2010): „Distribution based outlier detection for univariate data," *Statistics Netherlands*, The Hague/Heerlen, 10003

A.Vargha (2008): „Matematikai statisztika”, *Pólya kiadó*, Budapest

Az értekezés témaköréből írt, vagy megjelenés alatt álló tudományos közlemények

Magyar nyelvű közlemények

Hámori Gábor: „Fizetésképtelenség előrejelzése logit-modellel”. Bankszemle. 2001 / 1-2. p.65-87

Hámori Gábor: „Chaid-alapú döntési fák jellemzői”. Statisztikai Szemle 79. évfolyam 2001 / 8 p.703-710

Hámori Gábor-Csákány Tibor: „Szofisztikált kockázatkezelési módszerek egészségügyi alkalmazásokban/kórházi környezetben”, Kórház 2012/12 p.18-19

Hámori Gábor: „Érvényesség és korlátok az algoritmikus döntéshozatalban”. Gazdaság és Pénzügy 2015/4-es számába megjelenésre befogadásra került. p.1-11

Hámori Gábor:"Magyarázó változók kezelésének egyes kérdései regressziós modellezés során". Statisztikai Szemle 2016 januári számába befogadásra került. p.1-19

Idegen nyelven közlemények

Gábor Hámori: „Regression based classification models and expert judgement in predictive situations”, Regional and Business Studies 2015 Vol 7 No 1, 51-60

Egyéb

Best Papers, Global Spine Congress organized by AOSpine 2013, Hong Kong, April 4-6, 2013: Tibor Csákány, Gábor Hámori, P.P Varga: "Risk factors for surgical site infection following thoracolumbar spinal operations and a novel risk stratification model using predictive analytics".

Hámori Gábor: „Információszerzés nagyméretű adatbázisokból”, HVG Big Data konferencia, Budapest 2014. október. 28 (konferenciakötet megjelenés alatt)

Szakmai Életrajz

Hámori Gábor 1967. július 29-én született Budapesten. A budapesti Arany János Gimnáziumban érettségizett 1985-ben. A sorkatonai szolgálatot követően tanulmányait az Eötvös Lóránd Tudományegyetem Természettudományi Karán matematika-fizika szakon folytatta, ahol 1993-ban diplomázott. Még ugyanebben az évben felvételt nyert a Budapesti Közgazdaságtudományi és Államigazgatási Egyetemre (a mai Corvinus Egyetem), ahol Alkalmazott Statisztika- Gazdaságpolitika szakirányon szerzett diplomát 2000-ben. Az egyetemi tanulmányok mellett 1998-ig az Arany János, Táncsics Mihály és Pasaréti Gimnázium óraadó matematika és fizika tanáraként dolgozott. 1998-2000 között a Budapest Bank központi elemző részlegének vezető elemzőjeként folytatta pályafutását. 2001-2007 között az OTP Bank portfolió elemzési és scoring modellezési főosztályának vezetője. 2008-tól Alkalmazott Tudományért Kutatási és Fejlesztési alapítvány kurátora és a STAT4U adatbányászati kutatásvezetője. Több mint tíz éve meghívott előadó a Budapesti Corvinus Egyetem Statisztika és Befektetések tanszékén statisztikai esettanulmányok, statisztika II, és hitelkockázatok tárgyakban. Rendszeres lektori tevékenységet végez a Statisztikai Szemle, a Központi Statisztikai Hivatal havonta megjelenő tudományos folyóirata részére. Szűkebb szakterülete a többváltozós statisztikai modellek gyakorlati alkalmazása.

2002-ben felvételt nyert a Budapesti Műszaki és Gazdaságtudományi Egyetem Gazdálkodás- és Szervezéstudományok Doktori Iskolájába, ahol a doktori (PhD) képzés tanulmányi kötelezettségeit teljesítette, és erről végbizonyítványt (abszolutórium) kapott. Doktori tanulmányait 2012-ben a Kaposvári Egyetem Gazdálkodás- és Szervezéstudományok Doktori Iskolájában folytatta. 2013. februárban a doktori szigorlatot „summa cum laude” minősítéssel tette le. Angolul és olaszul beszél.