FLISZÁR VILMOS – KOVÁCS ESZTER – SZEPESVÁRY LÁSZLÓ –
SZÜLE BORBÁLA:

TÖBBVÁLTOZÓS ADATELEMZÉSI SZÁMÍTÁSOK

Feladatgyűjtemény

2016

Előszó

A helyesen alkalmazott többváltozós adatelemzési módszerekkel akár korábban nem ismert, érdekes szakmai összefüggések is felfedezhetők. A többváltozós adatelemzés tanulmányozásánál az elméleti (gyakran matematikai) és a gyakorlati számítási tudnivalók megismerése egyaránt fontos: a gyakorlati számítások eredményeinek helyes értelmezése a számítások elméleti hátterének pontos ismeretével valósulhat meg. Jelen írás elsősorban a gyakorlati számítások konkrét részletkérdéseivel foglalkozik és feltételezi, hogy az Olvasó az elemzési módszerek elméleti háttérét már ismeri.

A többváltozós adatelemzési számítások a gyakorlatban gyakran valamilyen program (illetve programcsomag) segítségével végezhetők el. Ebben az írásban a bemutatott elemzési módszerekkel kapcsolatos számításokat az IBM SPSS Statistics 20 programcsomag alkalmazásával szemléltetjük. Mivel a számítások elméleti háttere szorosan összefügg az eredmények értelmezésével, ezért a következőkben mindegyik fejezet az adott témához kapcsolódó rövid elméleti összefoglalóval indul, amelyet számítási példák megoldásának bemutatása követ. Az egyes fejezetekben a számolásokhoz alkalmazott adatok az IBM SPSS Statistics 20 programcsomag minta adatfile-jai („sample files") között találhatók. A néhány számolási feladat megoldásának leírása után mindegyik fejezetben további (önálló gyakorlásra alkalmas) feladatok is találhatók.

A többváltozós adatelemzés tanulása gyakran időigényes folyamat, mivel a tanulás során az elméleti és a gyakorlati tudnivalók együttes megismerésére van szükség. Általában e tanulási folyamatot segítheti a számítások rendszeres gyakorlása, amelyhez sok sikert kívánnak

a Szerzők

1. fejezet

LEÍRÓ STATISZTIKAI MUTATÓSZÁMOK ELEMZÉSE

A módszer rövid összefoglalása

Bármely elemzésnél a megvalósítható elemzési módszereket nagyban befolyásolja a rendelkezésre álló adatok köre. Egy gyakorlati probléma megoldásánál az adatbázis előállítása általában sokkal több időt vesz igénybe, mint magának az elemzési folyamatnak a megvalósítása. A leíró statisztikák a meglévő adatokról, illetve változókról szolgálnak nagyon hasznos információval az elemző számára, melyek a későbbi lépések meghatározásában is segítséget nyújthatnak.

Megoldási módszerek és az eredmények értelmezése

A gyakorló feladatok megoldásánál említett változók a bankloan.sav adatai között találhatók.

1. feladat:
Elemezze az „Age in years" („age") változó tulajdonságait leíró statisztikai mutatószámokkal, illetve grafikonokkal!

A feladat megoldása:

Az „age" változó arányskálán mért változó, mely hiteligénylő ügyfelek életkorát tartalmazza. Az elemzést a következő menüpontjának segítségével végezhetjük el:

Analyze → Descriptive Statistics → Descriptives

Ezen menüpont választása után az egyik eredmény a következő táblázat:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Age in years	850	36	20	56	35,03	8,041	64,665	,335	,084	-,658	,168
Valid N (listwise)	850										

A kapott táblázatból látható, hogy az adatbázisban 850 ügyfélről találhatunk információkat, és mind a 850 esetén rendelkezésünkre áll az ügyfél életkora. Ebből következően esetleges további elemzéseknél hiányzó adatok pótlására nincsen szükség.

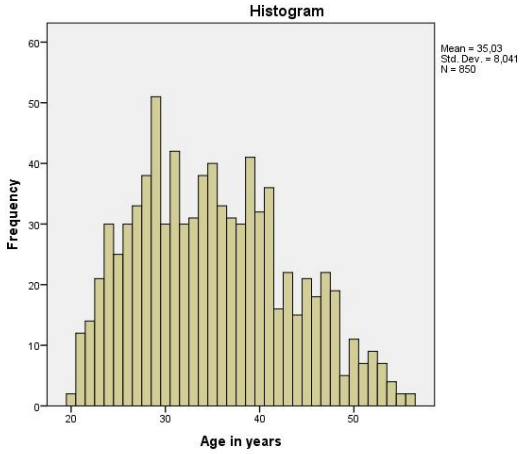
A legfiatalabb ügyfél 20, a legidősebb 56 éves, vagyis a változó terjedelme (range) 36. Az ügyfelek átlagos életkora 35,03 év, 8,041 év szórással. A variancia a szórásnégyzet. A relatív szórást (szórás/átlag) közvetlenül nem adja meg az SPSS, de a könnyen kiszámítható 0,23-as értéke belül marad a kritikus 2-es szinten. Kettő feletti relatív szórás esetén a változónkon belül a megfigyelések olyan mértékben szóródnának, hogy sok esetben a becslés stabilitását is befolyásolhatná az adott változó.

A változó eloszlásáról ad információt a csúcsosság (kurtosis) és a ferdeség (skewness). A 0,335-ös ferdeség érték enyhén jobbra elnyúló, a -0,658-as csúcsosság a normális eloszlás haranggörbéjénél lapultabb eloszlást jelez.

Ennek alátámasztásához állítsuk elő a változó hisztogramját. Ehhez válasszuk a következő menüpontot:

Analyze → Descriptive Statistics → Explore

Az „age” változót ezután a megfelelő módon kiválasztva a hisztogram az eredmények között található:

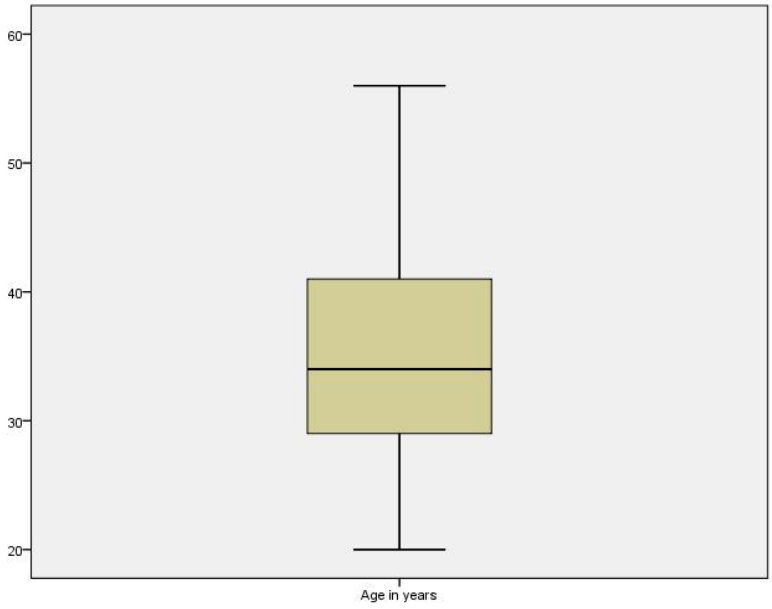


Az Explore menüpontban az SPSS egy úgynevezett stem&leaf ábrát is kirajzol, amely egy gyakorisági ábra, és felsorolja az egyes csoportokban előforduló értékeket. A megfigyelt érték utolsó számjegye a levél (leaf).

```
Age in years Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    14,00       2 .  00111111111111
    35,00       2 .  22222222222223333333333333333333
    55,00       2 .  4444444444444444444444444444555555555555555555555555
    63,00       2 .  666666666666666666666666666677777777777777777777777777777777
    89,00       2 .  88888888888888888888888888888888888888999999999999999999999999999999999999999999999999999
    72,00       3 .  000000000000000000000000000001111111111111111111111111111111111111
    61,00       3 .  22222222222222222222222222233333333333333333333333333333333
    78,00       3 .  444444444444444444444444444444444444455555555555555555555555555555555555555
    64,00       3 .  66666666666666666666666666666677777777777777777777777777777777
    71,00       3 .  88888888888888888888888888888899999999999999999999999999999999999999999
    68,00       4 .  00000000000000000000000000000011111111111111111111111111111111111
    38,00       4 .  22222222222222233333333333333333333
    36,00       4 .  444444444444445555555555555555555
    40,00       4 .  6666666666666666677777777777777777777
    24,00       4 .  888888888888888888999999
    18,00       5 .  000000000001111111
    16,00       5 .  2222222223333333
     6,00       5 .  444455
     2,00       5 .  66

Stem width:    10
Each leaf:      1 case(s)
```

Az ábrából látható, hogy összesen 14 fő 20-21 év közötti ügyfélnek folyósítottak hitelt, és közülük kettő 20 éves ügyfél található. 55 év felett összesen 4 ügyfél kapott csak hitelt. A további gyakorisági információkat analóg módon olvashatjuk ki, az értelmezésnél vegyük figyelembe az ábra feliratait.

Az Explore menüpontban szintén kérhetünk leíró statisztikai mutatókat a Descriptives menüponthoz hasonlóan, ezenkívül itt további mutatószámok is számíthatók. A nyesett átlag (trimmed mean) a megfigyelések alsó, illetve felső 5%-ának elhagyásával számított átlag. Esetünkben értéke 34,81, amely alig tér el a teljes sokaságra számított 35,03-as értéktől. Ez az eredmény arra is utalhat, hogy nincs sok extrém érték a változó esetében, és az extrém értékek hiányát mutatja a dobozdiagram is:



Az ábrából jól látható a leíró mutatószámok alapján is leolvasható koncentráció. A doboz belső része a felső és alsó kvartilisek (Q3-Q1) közötti távolságot, az interkvartilis terjedelmet - IQR (esetünkben 12) jeleníti meg, a dobozban lévő vonal a medián (34 év), amely minimálisan az átlag alatt van. Ez az eltérés is az enyhe ferdeségre utal. A doboz alatt illetve felett húzott vonalak hossza maximum 1,5-szer az interkvartilis terjedelem, amely természetesen rövidebb, ha elérjük a mintabeli maximumot vagy minimumot. Amennyiben egy megfigyelés a 1,5-szer IQR sávon kívül esik kilógó pontnak (outliernek), ha a 3-szor IQR sávon is kívül esik, akkor extrém értéknek tekintjük, amelyeket az SPSS o illetve * jelöléssel jelöl. A dobozdiagramból látható, hogy esetünkben nincsenek kilógó értékek.

A változó normalitásáról már a csúcsosság és a ferdeségi mutatók illetve a hisztogram is ad információt, az Explore menüpont azonban további két tesztet is számszerűsít, amelyek nullhipotézise a változó normalitása:

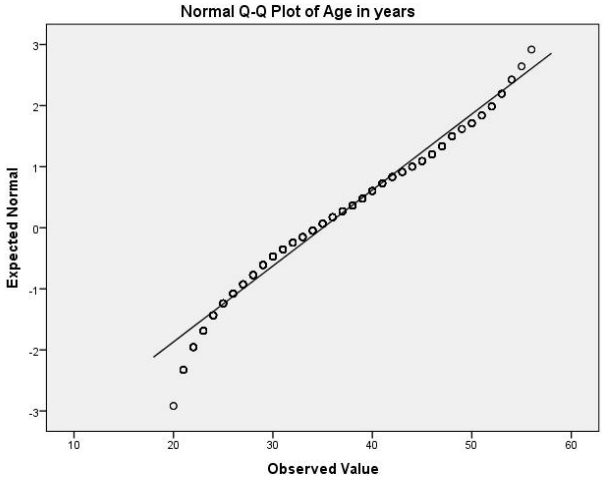
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age in years	,078	850	,000	,976	850	,000

a. Lilliefors Significance Correction

Az eredményekből jól látható, hogy mind a Kolmogorov-Smirnov, mind a Shapiro-Wilk teszt nullhipotézisét az összes szokásos szignifikanciaszinten elvetjük, vagyis az „age" változó nem tekinthető normális eloszlásúnak.

Az Explore-ban lehetőségünk nyílik még grafikus normalitás vizsgálat elvégzésére is, ehhez ún. Q-Q plotot kérhetünk. Ez nem más mint a kvantilisek ábrája, melynek vízszintes tengelyén az „age" változó értékei szerepelnek, a függőlegesen pedig egy olyan transzformációt alkalmazunk, amelynél normális eloszlás esetén az értékek a 45 fokos egyenes mentén helyezkednének el. Esetünkben jól látható, hogy a pontok a fiatalabb ügyfeleknél jelentősen eltértnek a 45 fokos egyenestől.



Összegzésként elmondható, hogy egy intervallum skálán mért változó SPSS-beli leíró statisztikai elemzéséhez mind a Descriptives, mind az Explore menüpontokat egyaránt fel kell használnunk, hiszen olyan elemeket is tartalmaznak, amelyekre a másik menüpont nem feltétlen ad lehetőséget. Emellett az elemzés során nem tértünk ki, de a későbbi elemzési módszertanok szempontjából fontos, hogy a Descriptives pontban lehetőség van a változó sztenderdizáltjának a mentésére.

2. feladat:

Elemezze a „Level of education" („ed") változó tulajdonságait leíró statisztikai mutatószámokkal, illetve grafikonokkal!

A feladat megoldása:

Az „ed” változó a hitelt felvevő ügyfelek végzettségét jeleníti meg végzettségi kategóriák szerint, vagyis egy ordinális skálán mért változó. Az elemzést a következő menüpont segítségével végezzük:

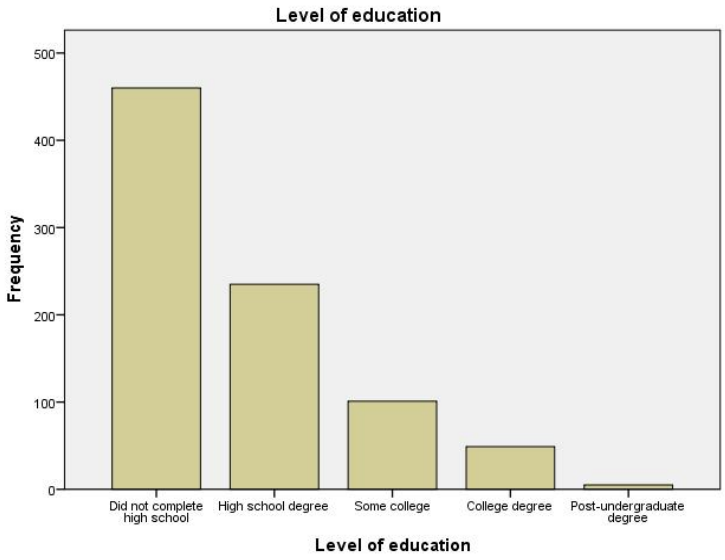
Analyze → Descriptive Statistics → Frequencies

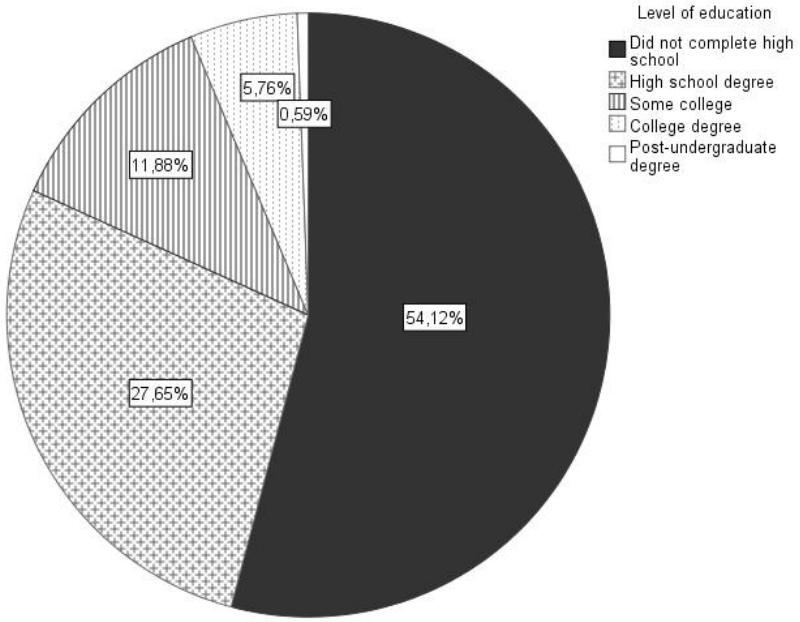
Az alap statisztikákból látható, hogy minden kategóriában található megfigyelés, vagyis az ügyfelek között a legalacsonyabb és legmagasabb végzettségűek is megtalálhatók. Az eloszlás módusza 1, vagyis a legtöbb ügyfél nem fejezte be a középiskolát. Ugyanakkor a gyakorisági tábla és az oszlopdiagram alapján jól látható, hogy a végzettségek eloszlása egy móduszú.

Level of education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Did not complete high school	460	54,1	54,1	54,1
	High school degree	235	27,6	27,6	81,8
	Some college	101	11,9	11,9	93,6
	College degree	49	5,8	5,8	99,4
	Post-undergraduate degree	5	,6	,6	100,0
	Total	850	100,0	100,0	

A megfigyelések 93,6%-a az első három kategóriában található, a legmagasabban mindösszesen 0,6%-uk, vagyis 5-en a 850-ből. Ez nagyon alacsony elemszám, amely minden valószínűség szerint nem hordoz többlet információt. Ezért a modellezés további lépései előtt érdemes végiggondolni az utolsó két kategória esetleges összevonását. A legnépesebb az első kategória, ahol a megfigyelések 54,1%-a található. A felsorolt információkat grafikusan kör és oszlopdiagramon is megjeleníthetjük. (Graphs\Chart Builder… menüpont)





Gyakorló feladatok

1. Hasonlítsa össze az „Age in years” („age”) változó és a Descriptives menüpontban létrehozott sztenderdizált („Zage”) változót a leíró statisztika eszközeivel! Mi tud mondani a két változó korrelációs együtthatójáról?

2. Hasonlítsa össze az „Age in years” („age”) változót és a centrálással létrehozott („C_age”) változót a leíró statisztika eszközeivel! Mi tud mondani a két változó korrelációs együtthatójáról?

3. Elemezze az „Age in years” („age”) változójának tulajdonságait úgy, hogy az elemzésnél válassza külön a default-os és a non-default-os ügyfeleket!

4. Definiálja a relatív szórást! Miért fontos az elemzésnél figyelembe venni? A kritikus értéke melyik „híres egyenlőtlenségből” származik?

5. Tegyük fel, hogy egy változó eloszlása egy móduszú és a változó átlaga nagyobb, mint a medián. Ekkor mit tud mondani az eloszlás tulajdonságairól?

6. Két móduszú eloszlás esetén mi mondható el a módusz, medián és átlag viszonyáról?

7. Amennyiben a vizsgált mintában a leíró statisztikai elemzés egyetlen egy extrém értéket jelöl, célszerű-e e kilógó értéket elhagyni a további elemzésből?

8. Egy kiemelten magas relatív szórású változó esetén milyen transzformációval lehetséges a relatív szórás csökkentése?

Irodalomjegyzék

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. Egy sztenderdizált változó szórása az eredeti (nem sztenderdizált) változó szórásához képest minden esetben

   a) nagyobb
   b) kisebb
   c) egyenlő
   d) egyik előző válasz sem helyes.

2. Három móduszú eloszlásnál minden esetben az átlag megegyezik

   a) a szórással
   b) az egyik módusszal
   c) a medián értékével
   d) egyik előző válasz sem helyes.

3. A sztenderdizálás hatására megváltozhat egy változó

   a) szórása
   b) a ferdeség értéke
   c) a csúcsosság értéke
   d) egyik előző válasz sem helyes.

4. A sztenderdizált változók

   a) átlaga és szórása egyenlő
   b) átlaga nagyobb mint a szórás
   c) átlaga kisebb mint a szórás
   d) átlaga és mediánja egyenlő.

2. fejezet

KERESZTTÁBLA ELEMZÉS

A módszer rövid összefoglalása

Kereszttábla-elemzéssel két nominális, két ordinális, vagy pedig egy ordinális és egy nominális mérési szintű változóra vonatkozóan lehet elemezni a változók közötti kapcsolat meglétét és a kapcsolat erősségét. A kereszttábla elemzésénél a nullhipotézis a két változó függetlensége. A függetlenség elvetésekor az asszociációs kapcsolat erősségét is lehet mérni különböző mutatószámokkal. A kapcsolaterősséget mérő mutatószámok közül általában szakmai megfontolások alapján lehet választani.

Megoldási módszerek és az eredmények értelmezése

A kereszttáblás elemzésnél is jelentősége van annak, hogy az egyes változók mérési szintje milyen (például ennek alapján lehet megfelelő kapcsolaterősségi mutatószámot választani). A következő feladatokban ezzel összefüggésben hangsúlyosan foglalkozunk a változók mérési szintjének témájával is.

A gyakorló feladatok megoldásánál említett változók az Employee data.sav adatai között találhatók.

1. feladat:
A „salary" változó alapján hozza létre a „salary_bin" változót úgy, hogy a „salary_bin" változó 4 kategóriája a „salary" változó egyes kvartiliseit jelölje!

A feladat megoldása:

Egy magas (arány) mérési szintű változónál elméletileg lehetőség van arra, hogy az értékeit (növekvő sorrendbe rendezés után) négy csoportba lehessen sorolni úgy, hogy minden csoport elemszáma azonos: az így létrehozott csoportokat kvartilisnek nevezzük. Az egyes kvartiliseket mutató változó létrehozása megoldható például a következő menüpont kiválasztásával:

Transform → Visual Binning …

A megjelenő ablakban a „salary" változót a „Variables to Bin" felirat alá elhelyezve és a „Continue" gombra kattintva újabb ablak jelenik meg, amelyben a kvartilisekhez kapcsolódó új változó létrehozásához a „Make Cutpoints …" gombnál választható a következő lehetőség:

Equal Percentiles Based on Scanned Cases → („Intervals – fill in either field" feliratnál) **Number of Cutpoints: 3**

A 3-as szám értéke azzal függ össze, hogy a kvartilisek száma 4. Az „Apply” gomb megnyomása után a „Binned Variable:” felirat után a „salary_bin” beírásával, majd az „OK” gomb megnyomásával létrejön az új, kvartilisekhez kapcsolódó kategóriás változó.

2. feladat:
Mennyi az elemek száma a „salary_bin” változó egyes kategóriáiban?

A feladat megoldása:

A kvartilisek egyenként elméletileg az összes adat egynegyedét tartalmazzák, gyakorlatilag azonban előfordulhat, hogy a kvartilisekben található elemek száma nem egyenlő. Ilyen eset előfordulhat például akkor, ha a megfigyelések száma (n) nem osztható néggyel, illetve ha a változó értékei között azonos értékek is vannak. A feladatban létrehozott „salary_bin” változó esetében a kategóriákba tartozó megfigyelések száma a következő menüpont választásával is számolható:

Analyze → Descriptive Statistics → Frequencies …

A változók közül a „salary_bin” választásával és a „Display frequency tables” lehetőség bejelölésével az „OK” gomb megnyomása után a következő eredmény számolható:

Current Salary (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	120	25,3	25,3	25,3
	2	117	24,7	24,7	50,0
	3	119	25,1	25,1	75,1
	4	118	24,9	24,9	100,0
	Total	474	100,0	100,0	

A „Frequency” feliratú oszlopban látható (bekarikázással is jelölt) értékek mutatják, hogy az egyes kategóriákban mennyi megfigyelés található (ezek az értékek ebben a példában nem egyenlők). Az eredmények alapján az is látszik, hogy a kategóriák jelölése a „salary_bin” változó esetében az 1, 2, 3 és 4 értékekkel történt.

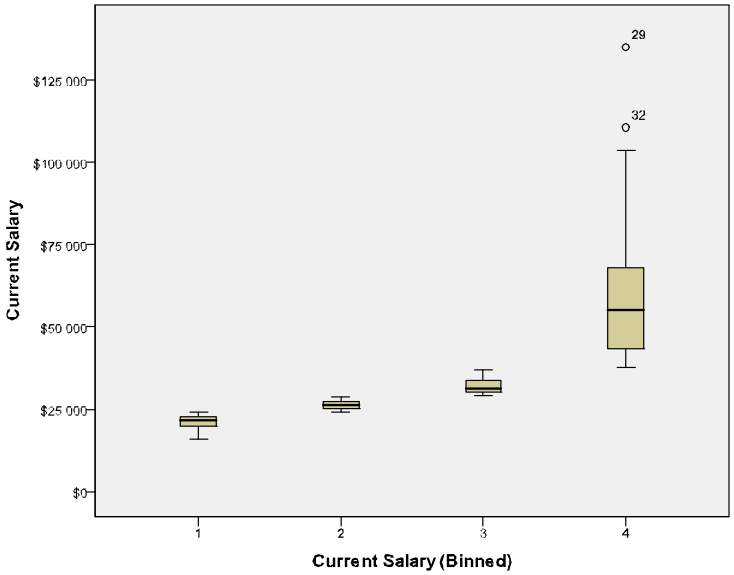
3. feladat:
Milyen mérési szintű a „salary_bin” változó?

A feladat megoldása:

A „salary_bin” változó ordinális mérési szintű, mivel a kvartilisek sorbarendezése jól értelmezhető feladatot jelent (matematikai értelemben is). Ez a „salary” és a „salary_bin” változókat közös dobozdiagramon ábrázolva is szemléltethető. Ez a dobozdiagram előállítható például a következő menüpont választásával:

Analyze → Descriptive Statistics → Explore …

A „salary” változót a „Dependent list:”, a „salary_bin” változót pedig a „Factor List:” feliratnál feltüntetve az „OK” gomb megnyomása után több eredmény között megtalálható a következő dobozdiagram is:



4. feladat:
A „salbegin” változó alapján hozza létre a „salbegin_bin” változót úgy, hogy a „salbegin_bin” változó 4 kategóriája a „salbegin” változó egyes kvartiliseit jelölje!

A feladat megoldása:

Az 1. feladat megoldásához hasonlóan oldható meg a feladat (az 1. feladatban említett „salary” változó helyett a „salbegin” változót alkalmazva a megoldásoknál).

5. feladat:
A „jobcat” és az 1. feladatban létrehozott „salary_bin” változó alapján végezzen kereszttáblás elemzést. Az elemzés eredményei alapján a két változó egymástól függetlennek tekinthető?

A feladat megoldása:

A kereszttáblás elemzést a következő menüpont választásával lehet végezni:

Analyze → Descriptive Statistics → Crosstabs …

A „Row(s)” feliratú részbe a „jobcat” változót, a „Column(s)” feliratú részbe a „salary_bin” változót helyezhetjük el (a változók természetesen felcserélve is elhelyezhetők lennének). A „Display clustered bar charts” lehetőséget bejelölve a két változó összefüggéséről grafikus megjelenítést is lehet kérni. A két változó függetlenségének teszteléséhez kapcsolódó Pearson-féle khi-négyzet próba tesztstatisztika értéke és az ezzel összefüggő empirikus szignifikancia-szint a

„Statistics…” gomb megnyomása után a „Chi-square” lehetőség bejelölésével számolható. Az eredményeket a „Continue” és ezután az „OK” gomb megnyomása után lehet megtekinteni.

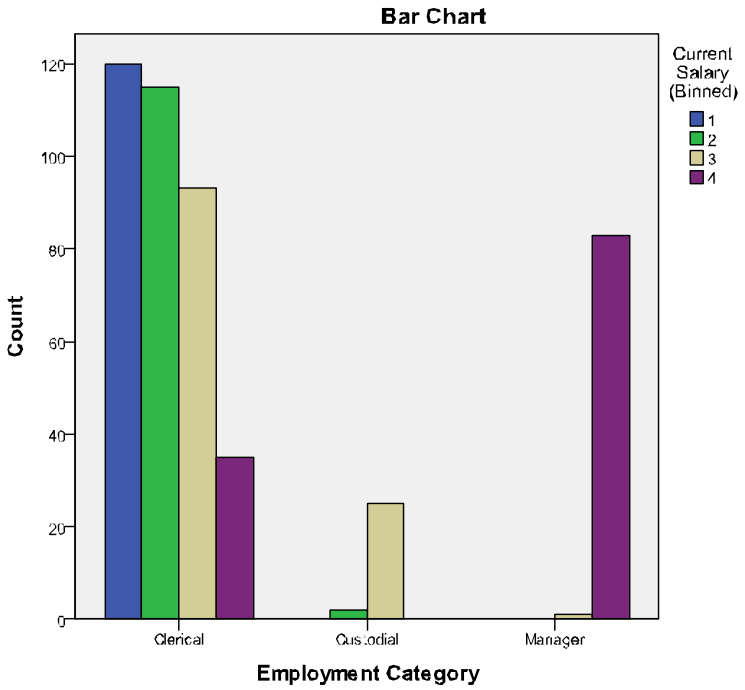
Ebben a feladatban a következő kereszttábla is megtalálható az eredmények között:

Employment Category * Current Salary (Binned) Crosstabulation

Count

		Current Salary (Binned)				Total
		1	2	3	4	
Employment Category	Clerical	120	115	93	35	363
	Custodial	0	2	25	0	27
	Manager	0	0	1	83	84
Total		120	117	119	118	474

A kereszttáblában a bekarikázással jelölt értékek a 2. feladat megoldásaként is számolhatók. A “Display clustered bar charts” lehetőség bejelölésének eredményeként a kereszttáblában szereplő adatokat (empirikus gyakorisági értékeket) szemléltető oszlopdiagram is megtalálható az eredmények között:



Tegyük fel, hogy a kereszttáblában az empirikus gyakorisági értékeket g_{ij} jelöli és m_i a sorokban szereplő értékek összegére, n_j pedig az oszlopokban szereplő értékek összegére utal (tegyük fel, hogy a sorok száma r, az oszlopok száma pedig c). Ebben az esetben az empirikus értékek alapján kiszámolható a

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(g_{ij}-\frac{m_i\cdot n_j}{n}\right)^2}{\frac{m_i\cdot n_j}{n}}$$

mutatószám (amelynek elméleti eloszlása a két változó függetlensége esetén χ^2 eloszlás, ahol a szabadságfok $(r-1)\cdot(c-1)$), amelynek alapján következtetni lehet a két változó függetlenségéhez kapcsolódó nullhipotézis elfogadására vagy elutasítására. Ennek a mutatószámnak az értéke elméletileg nulla akkor, ha két független változó szerepel az elemzésben. Az elméletileg khi-négyzet eloszlású mutatószámhoz kapcsolódóan empirikus szignifikancia-szint (p-érték) is számolható, amelyet a gyakorló feladat eredményei között a következő táblázatban a bekarikázással jelölt érték mutat:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	361,043[a]	6	,000
Likelihood Ratio	341,990	6	,000
Linear-by-Linear Association	193,616	1	,000
N of Valid Cases	474		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 6,66.

Mivel ebben a feladatban az empirikus szignifikanciaszint értéke alacsony (kisebb mint a gyakran alkalmazott 0,05), így a két változó függetlenségére vonatkozó nullhipotézist elutasíthatjuk, vagyis a két változó függetlensége elvethető.

6. feladat:

A „jobcat" és a 4. feladatban létrehozott „salbegin_bin" változó alapján végezzen kereszttáblás elemzést. Az elemzés eredményei alapján a két változó egymástól függetlennek tekinthető?

A feladat megoldása:

Az 5. feladat megoldásához hasonlóan számolható a khi-négyzet értéket és az ehhez kapcsolódó empirikus szignifikancia-szintet (p-értéket) is tartalmazó táblázat:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	302,359[a]	6	,000
Likelihood Ratio	294,130	6	,000
Linear-by-Linear Association	187,013	1	,000
N of Valid Cases	474		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,24.

Mivel a fenti táblázatban a (bekarikázással jelölt) empirikus szignifikancia-szint alacsonyabb, mint a gyakran alkalmazott 0,05, így a függetlenségre vonatkozó nullhipotézis elvethető. Ez az eredmény azt is jelenti, hogy lehet értelme valamilyen (megfelelő) mutatószám alkalmazásával a két változó közötti kapcsolaterősség mérésének is.

7. feladat:

A „salary_bin" vagy a „salbegin_bin" változónak erősebb a kapcsolata a „jobcat" változóval?

A feladat megoldása:

Ez a feladat megoldható úgy is, hogy az 5. és 6. feladatok eredményei között szereplő empirikus szignifikanciaszinteket, illetve az ezekhez tartozó khi-négyzet értékeket hasonlítjuk össze. Az 5. feladatban a „salary_bin" és a „jobcat" változó esetében a khi-négyzet érték 361,043 volt, míg a 6. feladatban a „salbegin_bin" és a „jobcat" változó esetében a khi-négyzet érték ennél alacsonyabb, mindössze 302,359 volt. Ezek alapján a „jobcat" változónak a „salary_bin" változóval erősebb a kapcsolata.

8. feladat:

Hogyan mérhető a „salary_bin" és a „jobcat" változók kapcsolatának erőssége?

A feladat megoldása:

Az 5. feladat eredménye alapján a két változó nem tekinthető egymástól függetlennek, tehát érdemes foglalkozni a kapcsolaterősség mérésével is. A 3. feladat eredménye alapján a „salary_bin" változó ordinális mérési szintű, míg a „jobcat" változó akár nominális mérési szintű változónak is tekinthető. Abban az esetben, ha a „jobcat" változóról feltételezzük, hogy nominális mérési szintű, a kapcsolaterősség méréséhez nem alkalmazhatók azok a mutatószámok, amelyek feltételezik, hogy mindkét elemzésben szereplő változó mérési szintje ordinális (ilyen például a Goodman-Kruskal gamma).

A kapcsolaterősség mutatószámait esetenként asszociációs mértékeknek is szokás nevezni. Az asszociációs mértékek között szimmetrikus és nem-szimmetrikus mutatószámok is találhatók (a nem-szimmetrikus mutatószámoknál feltételezhető

hogy az egyik változó a másikra hat és nem fordítva, míg a szimmetrikus mutatószámoknál ez nem feltételezhető).

A „jobcat" és a „salary_bin" változó esetében feltételezhető, hogy a „salary_bin" kategóriákat befolyásolhatja a „jobcat" változó, tehát az asszociációs mértékek közül nem-szimmetrikus mutatószámot érdemes választani az elemzéshez.

Az asszociációs mérték számolása a kereszttáblás elemzésnél a következő menüpontnál választható ki:

Analyze → Descriptive Statistics → Crosstabs …

A „Statistics…" gombnál a „Nominal" feliratnál kiválasztható például a „Lambda" lehetőség, majd a „Continue" és az „OK" gomb megnyomása után számos asszociációs mértékről található adat az eredmények között (amit a következő táblázat is mutat):

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,335	,033	8,452	,000
		Employment Category Dependent	,432	,074	4,513	,000
		Current Salary (Binned) Dependent	,305	,024	11,827	,000
	Goodman and Kruskal tau	Employment Category Dependent	,475	,042		,000[c]
		Current Salary (Binned) Dependent	,253	,011		,000[c]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

Az előzőekben említett megfontolások alapján a lambda mutatószám esetében az előző táblázatban szereplő értékek közül tehát a bekarikázott érték (0,305) értelmezése lehet releváns: ez az érték a két változó esetében viszonylag gyenge (közepesen erősnek is nevezhető) kapcsolatra utal.

Érdemes még megemlíteni, hogy ha két ordinális változó szerepel az elemzésben, akkor az asszociációs mértékek előjele a változók közötti kapcsolat irányára is utal (vagyis az ordinális mutatószámokra vonatkozó asszociációs mutatószámok előjele negatív is lehet).

Gyakorló feladatok

1. Állítson elő 10 (megközelítőleg) azonos elemszámú csoportot Visual Binning alkalmazásával a „Current salary" változó alapján! A csoporttagságot mutató új változó neve legyen „csoportok"!

2. Kereszttáblás elemzéssel tesztelje, hogy a „csoportok" változó és a „jobcat" változó függetlennek tekinthető-e!

3. Alkalmazható lenne a 2. feladatban a kapcsolaterősség mérésére a „kappa” mutatószám?

4. Lehetséges lenne a 2. feladatban szereplő változók esetében nem-szimmetrikus asszociációs mértéket választani?

5. Milyen összefüggés van a kereszttáblás elemzés nullhipotéziséhez kapcsolódó khi-négyzet érték és a kontingencia együttható értéke között?

Irodalomjegyzék

Hajdu Ottó [2003]: *Többváltozós statisztikai számítások*
Központi Statisztikai Hivatal

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. A kereszttáblás elemzésben a kereszttábla sorainak száma

   a) nagyobb mint az oszlopok száma
   b) az oszlopok számával egyezik meg
   c) független az oszlopok számától
   d) egyik előző válasz sem helyes.

2. Mennyi változó szerepelhet egyidejűleg a kereszttáblás elemzésben?

   a) egy
   b) kettő
   c) mindkét előző válasz helyes
   d) egyik előző válasz sem helyes.

3. Bináris változók esetében végzett kereszttáblás elemzésben a változók függetlenségére vonatkozó nullhipotézisnél a tesztstatisztika eloszlása

   a) normális eloszlás
   b) khi-négyzet eloszlás
   c) F-eloszlás
   d) t-eloszlás.

4. A kereszttáblás elemzésben a kappa mutatószám

   a) értéke függ a kereszttáblában szereplő mindegyik értéktől
   b) négyzetes tábla esetén számolható
   c) értéke nulla ha bináris változók szerepelnek az elemzésben
   d) egyik előző válasz sem helyes.

5. A kereszttáblás elemzésben a Cramer-V mutatószám és a kontingencia együttható értéke

   a) elméletileg egyaránt maximum egységnyi lehet
   b) szorzata egységnyi
   c) összege egységnyi
   d) egyik előző válasz sem helyes.

3. fejezet

KLASZTERELEMZÉS

A módszer rövid összefoglalása

A klaszterelemzés olyan osztályozó eljárás, amellyel adattömböket – megfigyelési egységeket és változókat egyaránt – tudunk viszonylag homogén csoportokba sorolni az elemzésbe bevont változók alapján. A folyamat akkor sikeres, ha az egységek hasonlítanak csoporttársaikhoz, azonban eltérnek a más csoportba tartozó elemektől. A klaszterezés olyan felügyelet nélküli (unsupervised) osztályozást jelent, amelyben nincsenek előre definiált osztályok. A csoportképzés alapja a sokaság elemeinek elhelyezkedése a *p* dimenziós térben, amikor a sokaság egy-egy eleme a tér egy-egy pontja.

A klaszterezés többféle módszer és konkrét eljárás összefoglaló neve. A gyakorlati jelentőséggel és számítógépes kidolgozottsággal bíró klaszterezési módszereknek két fő csoportja van: hierarchikus és nemhierarchikus klaszterezés.
A hierarchikus klaszterezés elsősorban feltáró elemzésre szolgál, mivel nincs feltevésünk arra, hogy a minta hány klaszterre tagolódik. Összevonó (agglomeratív - gyakorlatban ezt használjuk) és a felosztó (divizív) szemléletben végezhető el.
A nemhierarchikus klaszterezés szakmai megfontolások alapján előre adott *k* számú osztályra bontja a mintát; megerősítő elemzésre szolgál.

A klaszterező eljárásokat széleskörűen alkalmazzák ismeretlen adatstruktúrák feltárására. A klaszterezés eredményeinek felhasználása során az alábbi megszorításokat szükséges figyelembe venni:

- elsősorban feltáró technikaként használható;
- nincs egyetlen legjobb megoldás;
- minden esetben létrehoz az eljárás klasztereket, függetlenül attól, hogy azok ténylegesen léteznek-e az adatokban;
- a megoldások függnek a bevont változóktól.

Megoldási módszerek és az eredmények értelmezése

A gyakorló feladatokat a bankloan.sav adatain végezzük el.

NEMHIERARCHIKUS KLASZTEREZÉS

1. feladat:
a) Kérjen nemhierarchikus klaszterelemzést 5 csoporttal a következő változókra:

'Age in years [age]'
'Years of current employer [employ]'
'Years of current address [address]'
'Household income in thousands [income]'
'Debt to income x100 [debtinc]'
'Credit card debt in thousands [creddebt]'
'Other debt in thousands [othdebt]'

Tekintse át az eredményeket!

b) Ismételje meg a nemhierarchikus klaszterelemzést azzal a módosítással, hogy a hitelkártya adósságot nem ezres nagyságrendben, hanem pontos összegben méri (azaz hozzon létre egy új változót a 'creddebt' változót ezerrel felszorozva)! Hasonlítsa össze az eredményeket és vonja le a következtetéseket!

Megjegyzés a feladathoz: a klaszterelemzésnél az esetek többségében szükséges sztenderdizálni a bevont változókat. Ennél a feladatnál azért nem a sztenderdizált változókat használjuk, hogy összevetve a további feladatok eredményeivel szemléltessük a sztenderdizálás hiányának hatását.

A feladat megoldása:

a) A nemhierarchikus klaszterelemzés a következő menüpont kiválasztásával érhető el:

Analyze → Classify → K-Means Cluster...

A 'Variables' dobozba áthelyezzük a megadott változókat. A kért klaszterszámot a 'Number of Clusters' melletti dobozban adhatjuk meg. A 'Method' dobozban célszerű az 'Iterate and classify' lehetőséget választani.[1]
A jobboldali 'Options' menüpontban egyelőre csak az egyedi klaszterközéppontokra van szükségünk ('Initial cluster centers').

A kapott klaszterek elemszámát a 'Number of Cases in each Cluster' táblázatból olvashatjuk ki. Ebből látható, hogy a kapott klaszterek elemszámai nagy különbséget mutatnak: egy klaszterbe került 545 ügyfél (az összes ügyfél 64,1%-a), további 229 ügyfél (26,9%) egy másik klaszterbe. Az ügyfelek megmaradt 9%-a három klaszterben oszlik szét. Célszerű lenne csökkenteni a klaszterek számát.
A 'Final Cluster Centers' táblázat segítségével jellemezhetjük az egyes klasztereket a klaszterközéppontjuk (centroidjuk) segítségével. Mivel nem sztenderdizált adatokból dolgoztunk, az eredeti mértékegység szerint kapjuk meg a centroidok elhelyezkedését a változók által képzett hét dimenziós térben.

b) Az ismételt futtatáshoz szükséges új változót a **Transform → Compute Variable** menüpontban állíthatjuk elő. A baloldali 'Target Variable' dobozban adjuk meg az új változó nevét (pl. creddebt2), a jobboldali 'Numeric Expression' dobozban pedig a szükséges transzformációt (pl. creddebt*1000). Az előző futtatást ismételjük meg azzal a módosítással, hogy a 'creddebt' változót a 'creddebt2' változóra cseréljük.

A két klaszterelemzés outputját összehasonlítva megállapíthatjuk, hogy a kezdeti és a végleges klaszterközéppontok ('Initial Cluster Center', 'Final Cluster Center') különböznek egymástól, nemcsak a hitelkártya adósság nagyságában, hanem a többi változó szerint is. Az egyes klaszterekbe sorolt ügyfelek száma változik. Továbbra is két nagyobb klaszterünk van, illetve három kisebb, azok számossága azonban különbözik az a) részben kapott klaszterekétől.

[1] Az 'Iterate and classify' esetén az iteráció során az eljárás a besorolt elemekre új klaszterközéppontot számol, majd újra besorol, míg 'Classify only' esetén csak a kezdeti középpontokhoz osztja szét a mintát.

Number of Cases in each Cluster

Cluster	1	64,000
	2	545,000
	3	11,000
	4	1,000
	5	229,000
Valid		850,000
Missing		,000

K-középpontú klaszter creddebt változóval

Number of Cases in each Cluster

Cluster	1	1,000
	2	584,000
	3	8,000
	4	201,000
	5	56,000
Valid		850,000
Missing		,000

creddebt2 változóval

Az eredmények alapján megállapíthatjuk, hogy a természetes mértékegységben mért változók nem tesznek eleget a klaszterezés egyik követelményének, nem invariánsak a lineáris transzformációra. Ennek kiküszöbölésére szükséges a változók sztenderdizálása. Általános megállapítás, hogy a klaszterezés során a különböző mértékegységben mért változók esetén a sztenderdizált formájukat célszerű használni, hogy az eredmény ne függjön a változók mértékegységétől és nagyságrendjétől.

2. feladat:

a) Sztenderdizálja az 1. feladatban használt változókat, majd a sztenderdizált változókra ismételje meg a nemhierarchikus klaszterelemzést 5 csoporttal! Milyen változás történt az 1. feladat eredményeihez képest?

b) Értelmezze a kapott eredményeket! Jellemezze az egyes klasztercsoportokat a klaszterközéppontok alapján! Próbálja megnevezni az egyes csoportokat a fő jellemzőik alapján!

c) Elemezze a bevont változók szerepét a klaszterezésben! Melyik változó játssza a legfontosabb szerepet a csoportképzésben? Ki lehet-e hagyni valamelyik változót az elemzésből?

d) Melyik megfigyelés van a saját klaszterközéppontjától a legmesszebb?

e) Hány klasztert érdemes kérni a 'könyökszabály' alapján?

A feladat megoldása:

a) A változók sztenderdizálást többféle módon elvégezhetjük SPSS-ben. Az egyik lehetőség az **Analyze → Descriptive Statistics → Descriptives...** parancs alkalmazása. A 'Variable(s)' dobozba áthelyezzük a bevonandó változókat, majd kérjük a 'Save standardized values as variables' lehetőséget.
Az 1. feladatban megismert módon, a létrejött változókkal futtassuk le a K-középpontú klaszterelemzést. A feladat további pontjaihoz kérjük az 'Options...' menüpontban az ANOVA-táblát, illetve a 'Save...' menüpontban a klaszterazonosítók és klaszterközéppontoktól mért távolságok mentését is.

Az 1. feladat klaszterelemzéseihez képest megváltozik a kezdő és a végső klaszterközéppontok nagysága, a sztenderd normális eloszlás szerinti, mértékegység nélküli értéküket kapjuk meg. Az egyes csoportok jellemzésénél a viszonyítási pont az egyes változók sztenderd normális eloszlás szerinti átlagos értéke, a nulla.

Megváltozik az egyes klaszterek nagysága. Továbbra is van két nagyobb klaszterünk, ezek azonban a korábbihoz képest csak az összes ügyfél 70,7%-át fedik le.

b) Az 1. feladat klasztereihez képest kiegyensúlyozottabb a struktúra, bár így is jelentős a különbség az egyes klaszterek számossága között.
A második legnagyobb méretű az 1-es klaszter (250 fő, 29,4%), viszonylag idősebb, de nem a legidősebb csoport, akik hosszabb ideje laknak, dolgoznak azonos helyen. A csoport jövedelme átlag körüli (kis mértékben alatta van, mivel a 'Final Cluster Centers' táblázatban a 'Household income in thousands' változó értéke -0.06235, ami közel van az átlagos értéket kifejező 0-hoz), relatív és abszolút eladósodottságuk átlag alatt marad. Ez alapján a megfontolt, kevésbé eladósodott idősebb csoport elnevezést adhatnánk nekik.
A többi csoport részletes jellemzését a 'Final Cluster' táblázat alapján az Olvasóra bízzuk. A továbbiakban az egyes csoportok általunk adott rövid elnevezését adjuk meg:

- 2-es klaszter (97 fő): idős, immobil, kis mértékben eladósodott ügyfelek csoportja
- 3-as klaszter (14 fő): idős, jól kereső, magas hitelállománnyal rendelkezők kis csoportja
- 4-es klaszter (138 fő): átlagos életkorú, átlagnál jobban eladósodott ügyfelek csoportja
- 5-ös klaszter (351 fő): fiatal, mobil, alacsony jövedelmű, az átlaghoz képes kevésbé eladósodott ügyfelek csoportja

Final Cluster Centers

	Cluster				
	1	2	3	4	5
Zscore: Age in years	,60121	1,19503	1,39979	,01166	-,81888
Zscore: Years with current employer	,20510	1,46339	2,07690	-,00758	-,63035
Zscore: Years at current address	,60917	,93140	,81628	-,19580	-,64686
Zscore: Household income in thousands	-,06235	1,43353	4,31900	-,04365	-,50685
Zscore: Debt to income ratio (x100)	-,46159	-,02570	,98963	1,53331	-,30644
Zscore: Credit card debt in thousands	-,29626	,60180	4,77982	,72578	-,43129
Zscore: Other debt in thousands	-,34967	,94669	3,87222	,78734	-,47657

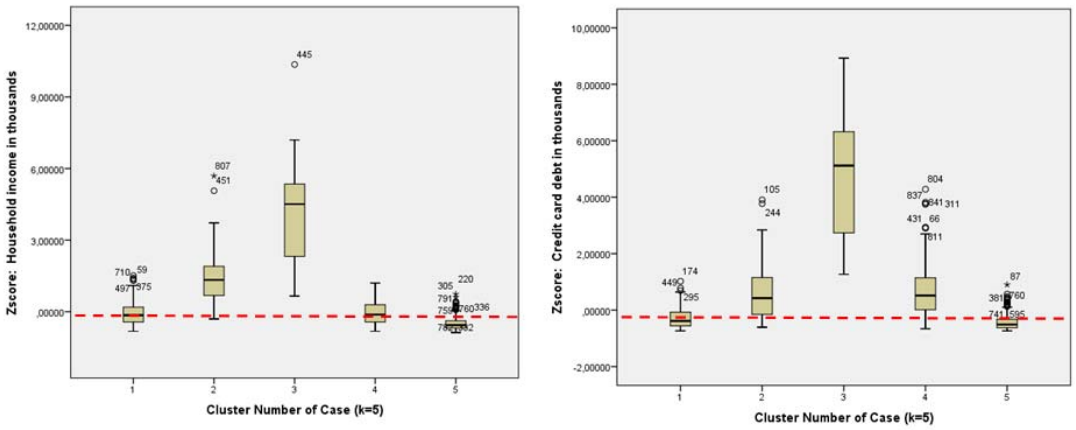
c) A bevont változók szerepét az ANOVA tábla segítségével elemezhetjük. Azonban fontos megemlíteni (erre az SPSS is figyelmeztet), hogy az ANOVA elemzés feltételei (a bevont változók minden csoportban normális eloszlásúak, az egyes csoportokban a szórásúk azonos) a legtöbb esetben nem teljesülnek. A legfontosabb szerepet a csoportképzésben a 'Household income in thousands [income]' változó játssza, mivel ez a változó rendelkezik a legmagasabb F értékkel (392,414). Emellett a 'Credit card debt in thousand [creddebt]' változó hatása is jelentős (F értéke 325,582).

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore: Age in years	122,927	4	,423	845	290,723	,000
Zscore: Years with current employer	104,526	4	,510	845	204,979	,000
Zscore: Years at current address	84,602	4	,604	845	140,010	,000
Zscore: Household income in thousands	137,974	4	,352	845	392,414	,000
Zscore: Debt to income ratio (x100)	106,112	4	,502	845	211,197	,000
Zscore: Credit card debt in thousands	128,727	4	,395	845	325,582	,000
Zscore: Other debt in thousands	123,171	4	,422	845	292,096	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

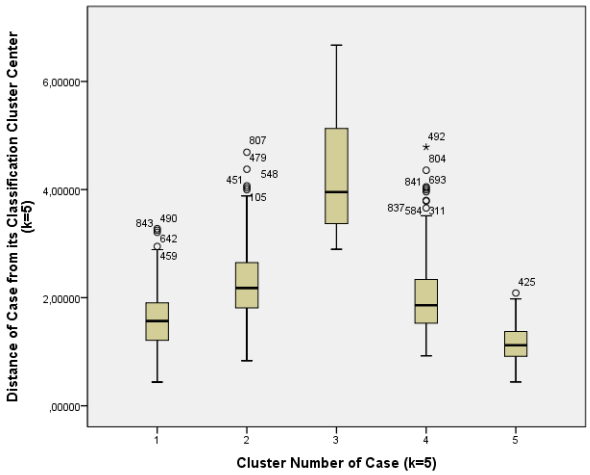
Érdemes klaszterenként boxplottal ábrázolni ezen változók értékét[2]. Az ábrák alapján megállapíthatjuk, hogy a legmagasabb F értékek ellenére az 'income' változó esetén az 1-es és a 4-es klaszter, a 'creddebt' változó esetén az 1-es és az 5-ös klaszter sztenderdizált értékei hasonlók, a másik három klaszter esetén jelentősen különböznek a változók értékei. Azaz a magas F érték nem jelenti egyértelműen valamennyi klaszter jelentős eltérését egymástól, de összességében biztos van közöttük kettő, amely különbözik.
Mivel valamennyi változóhoz tartozó F érték szignifikáns bármely szignifikancia szint mellett (lásd ANOVA tábla utolsó 'Sig' oszlopa, mindegyik változóhoz tartozó $p < 0{,}001$), egyik változó kihagyása sem indokolt.



d) A beállítások során kértük a klaszterközépponttól való távolság mentését. Az új változó értékeit leíró statisztikákkal vizsgálva megállapíthatjuk, hogy a maximális távolság a 445-ös megfigyeléshez tartozik (távolság értéke 6,67), amely a 3-as klaszterbe sorolódott be.
Érdemes boxplottal[3] ábrázolni klaszterenként a klaszterközéppontoktól vett távolságértékeket. Így megállapíthatjuk, hogy a legnagyobb klaszterközépponttól vett távolságok - nem meglepő módon - a 3-as klaszternél fordulnak elő, amely a 14 főből álló extrém ügyfélkört tartalmazza. A legkisebb középponttól vett távolságok az 1-es

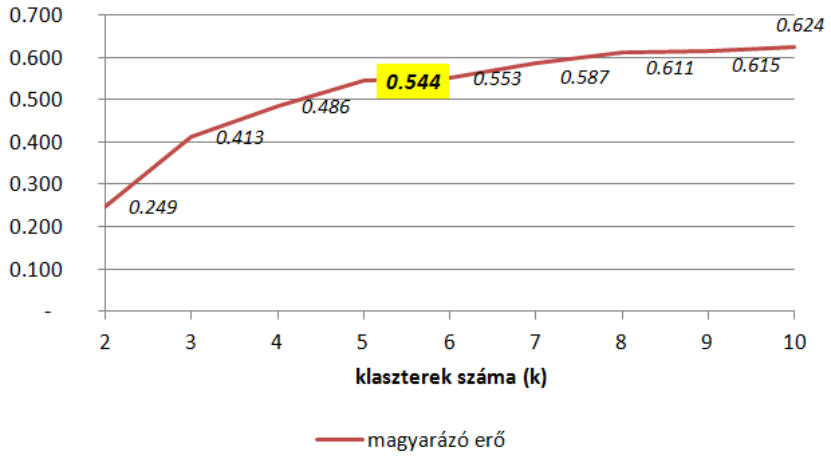
[2] Legegyszerűbb az **Analyze → Descriptive Statistics → Explore...** menüponton belül a 'Dependent List' dobozba áthelyezni a sztenderdizált 'income' és 'creddebt' változókat, a Factor list-hez behúzni az elmentett klaszterazonosítókat, a 'Plots' menüpontban csak a boxplotot kérni.
[3] Az **Analyze → Descriptive Statistics → Explore...** menüpontban tudjuk megrajzoltatni.

és az 5-ös, legnagyobb elemszámú klasztereknél tapasztalhatók, azaz azok viszonylag homogén csoportokat alkotnak.



e) Mivel a klaszterek száma nem haladhatja meg a $\sqrt{n/2}$, ezért a klaszterek maximális száma 20 lehet a példánkban. Sajnos a klaszterkönyök meghatározásához nincs beépített opció az SPSS-ben, saját magunknak kell meghatározni. Ehhez először szükséges a K-középpontú klaszterezést a feladat a) pontjának megfelelő beállításokkal, de különböző, eddig nem használt klaszterszám beállítások (k=2, 3, 7, 8, 9, ..., 20) mellett ismételten lefuttatni és a klaszterazonosítókat menteni. Figyeljünk arra, hogy azonosítani tudjuk, melyik új változó melyik klaszterszámú futtatáshoz tartozik.
Ezt követően valamennyi klaszterszám (k=2, 3, ..., 20) mellett egyutas ANOVA elemzést végzünk, amit az alábbi menüpontban érhetünk el:
Analyze → Compare means → One-way ANOVA...

A 'Dependent List' dobozba helyezzük át a feladat a) pontjában megadott változók sztenderdizált változatát, a 'Factor' dobozba helyezzük át a k=2 futtatás eredményeként kapott 'Cluster Number of Case' változót. A kimenet ANOVA tábláját másoljuk át például egy excel fájlba. Ott adjuk össze a Sum of Squares oszlopokban a változókhoz tartozó külső ('Between groups') eltérés négyzetösszegeket, majd a teljes ('Total') eltérés négyzetösszegeket is, a kapott két összeget pedig osszuk el egymással (külső/teljes). Így kapjuk meg az adott klaszterszám melletti magyarázóerőt. Ismételjük ezt meg valamennyi lehetséges klaszterszám mellett.
Érdemes a különböző klaszterszámok mellett kapott magyarázóerő értékeket ábrázolni. A klaszterkönyök, így az optimális klaszterszám ott lesz, ahol a magyarázóerő növekedése a legnagyobb. Jelen esetben öt klasztert érdemes használni a 'könyökszabály' alapján, hiszen ennél magasabb klaszterszámok esetén lassul a magyarázóerő növekedési üteme.



HIERARCHIKUS KLASZTEREZÉS

3. feladat:

a) Készítsen hierarchikus klaszterelemzést a főiskolai végzettségű ('ed' változó értéke 4 ['College degree']) ügyfelekre! A Ward eljárásban négyzetes euklideszi távolságot és az alábbi változók sztenderdizált formáját használja:

'Age in years [age]'
'Years of current employer [employ]'
'Years of current address [address]'
'Household income in thousands [income]'
'Debt to income x100 [debtinc]'
'Credit card debt in thousands [creddebt]'
'Other debt in thousands [othdebt]'

Tekintse át a kapott kimeneteket!

b) A dendrogram alapján hány csoportra lehetne bontani a főiskolai végzettségű ügyfeleket?

c) Melyik két főiskolai végzettségű ügyfél van legközelebb egymáshoz? Melyik két ügyfél van legtávolabb egymástól? Mekkorák a távolságok?

d) Ismételten futtassa le a hierarchikus klaszterelemzést azzal a módosítással, hogy a Ward eljárás helyett a 'between-groups linkage' módszert[4] használja! Ebben az esetben a főiskolai végzettségű ügyfeleket hány csoportra lenne érdemes bontani? Hogyan változik meg az ügyfelek csoportosulása? Hasonlítsa össze a dendrogramokat, illetve a távolságokat! Kik lesznek a legközelebbi, legtávolabbi ügyfelek? Mekkora közöttük a távolság?

A feladat megoldása:

a) Mivel a feladat a főiskolai végzettségű ügyfelekre vonatkozik, a **Data → Select Cases**... menüpontban végezzük el az ügyfelek szűrését. Az 'If condition is satisfied' lehetőséget kijelöljük, majd az 'If...' gombra kattintva adjuk meg a szűrési feltételt (például 'ed=4') a jobb felső üres dobozban.

Ezt követően a szűrt adattáblára végezzük el a nemhierarchikus klaszterelemzést, ami a következő menüpont kiválasztásával érhető el:

Analyze → Classify → Hierarchical Cluster...

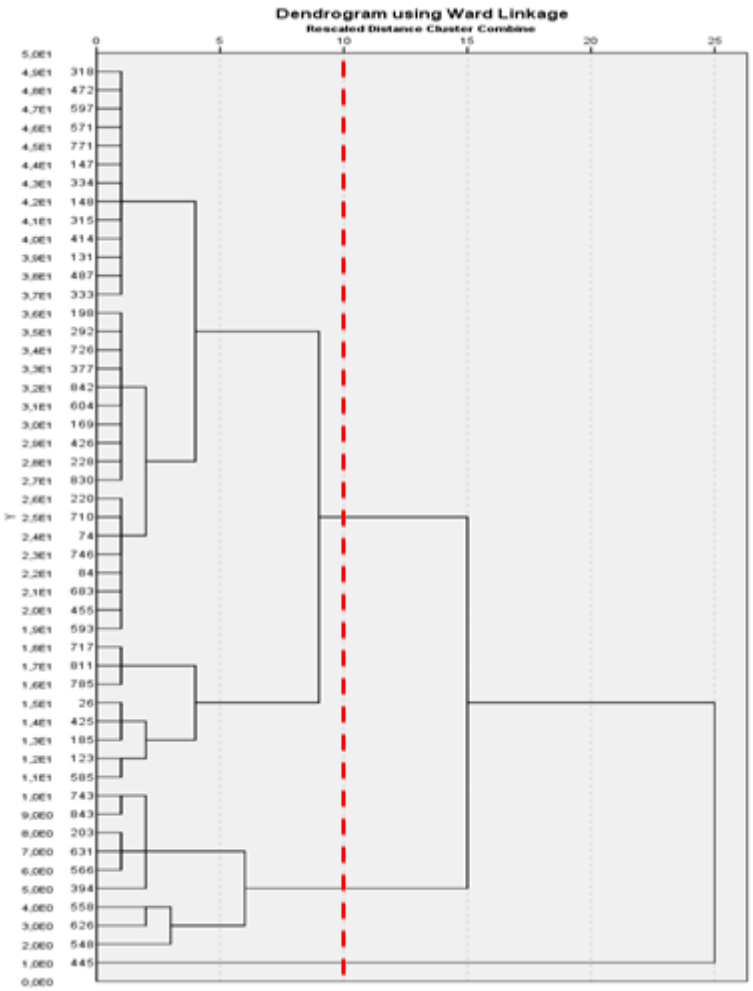
[4] Figyeljük arra, hogy az SPSS programban az egyes összevonási módszerek megnevezése különbözik a hierarchikus klaszter 'Method' beállításában és a kimenetekben.

A 'Variables(s)' dobozba helyezzük át a változókat. Itt választhatunk, hogy vagy az eredeti változókat használjuk és a 'Method' menüponton belül kérjük a sztenderdizálást, vagy rögtön a sztenderdizált változókat. A 'Cluster' dobozban, mivel most az ügyfeleket akarjuk klaszterezni, a 'Cases' opciót válasszuk, a 'Display' dobozban pedig a 'Statistics' és 'Plots' lehetőségeket is kérjük. A jobboldali menüpontok közül 'Statistics...'-ban válasszuk az 'Agglomeration schedule' és 'Proximity matrix' opciókat, a 'Plots...'-ban elegendő a 'Dendrogram'-ot kérni (többi lehetőséget nem szükséges módosítani). A 'Method...'-ban a 'Cluster Method' doboznál válasszuk a kért 'Ward's method'-ot, a 'Measure' blokkon belül pedig az 'Interval' beállítást kérjük a 'Squared Euclidean distance' távolsági mértékkel. A 'Method' parancson belül a 'Transform Values' dobozban tudjuk beállítani, hogy kérünk-e, illetve milyen transzformációt. Ha az eredeti változók használata mellett döntöttünk, itt a 'Z scores' és a 'By variable' választásával állíthatjuk be, hogy a sztenderdizált változókkal dolgozzunk a klaszterelemzésben.

A kimeneteken belül a 'Proximity Matrix' tartalmazza az egyes ügyfelek egymástól mért távolságát négyzetes euklideszi távolságmértékben meghatározva. A mátrix diagonális elemei az adott megfigyelés vagy változó (jelen esetben ügyfél) önmagától mért távolságát mutatják a kért távolságmérték szerint, így minden esetben 0 értéket vesz fel.
Az 'Agglomeration Schedule' ismerteti lépésről lépésre az egyes elemek (ügyfelek) összekapcsolódását (mivel 49 főiskolai végzettségű ügyfél van, így 48 lépésben vonja őket össze). A hierarchikus elvnek megfelelően kezdetben minden elem (ügyfél) külön klasztert alkot, majd lépésenként egy összevonás történik úgy, hogy az adott módszer szerinti két legközelebbi csoportot vonja össze az eljárás. A lépések addig ismétlődnek, míg a végére egy csoporttá áll össze az alaphalmaz. A táblázatból kiolvashatjuk, hogy az adott lépésben melyik két elem (ügyfél) kapcsolódik össze, mekkora közöttük a távolság, az egyes ügyfelek melyik klaszterben jelennek meg először, illetve melyik következő lépésben tűnnek fel ismét ezek az ügyfelek. Az alkalmazott Ward-féle eljárás azokat az elemeket (ügyfeleket) vonja össze, melyeknél az összevonás során a legkisebb lesz a belső (klaszteren belüli) szórásnégyzet növekedése.
A dendrogram speciális szerkezetben, két dimenzióban ábrázolja az ügyfelek összekapcsolódását. Az egyik tengelyen az összevont ügyfeleket látjuk, a másikon pedig azt a távolságértéket, amelynél az összevonás megtörtént. Kezdetben (0 távolsági szinten) minden megfigyelés önmagában van, a végén (SPSS-ben 25 maximális távolságértékre átskálázva) már minden ügyfél egyetlen csoportba kerül. A dendrogram segít az ügyfelek csoportjainak felderítésében. Elemzői szokás a 40%-os távolságszint (10-es rescaled distance) alatti csoportok számát leolvasni, és ezt elmenteni. Így két összevonó eljárás eredménye kereszttáblában is összevethető. Fontos azonban megjegyezni, hogy ebben az esetben nem kapunk végleges választ arra a kérdésre, hogy hány csoportba sorolható a vizsgált adathalmaz. A struktúrafeltárás ezen eljárása csak exploratív célra alkalmas, az ábra alapján hipotézis fogalmazható meg a mintabeli csoportok számára.

b) A 40%-os távolságszintnél három klaszter látható a dendrogramon, azonban a 3. klaszter csupán egy ügyfélből (445-ös) áll, aki legutoljára kapcsolódik a többi ügyfél által alkotott klaszterhez. Ez az ügyfél extrém változóértékkel rendelkezik, akit érdemes kizárni a további elemzésből. Ez a példa is rámutat arra, hogy a dendrogram

hatékonyan segíti az extrém értékek feltárását, hiszen a magas távolság szinten és/vagy az összekapcsolódás későbbi szakaszában látható megfigyelések egyedi jellege szembetűnő.
Összességében 445-ös ügyfél kivételével 2 klaszterbe érdemes sorolni a vizsgált ügyfeleket.



c) Az egymáshoz legközelebbi és a legtávolabbi ügyfeleket, illetve azok távolságát a 'Proximity Matrix' táblázatból olvashatjuk ki. A távolságmátrix alapján a legközelebbi ügyfelek a 318. és a 472. (köztük lévő négyzetes euklideszi távolság 0,070), a legtávolabbiak a 318. és a 445. ügyfél (köztük lévő négyzetes euklideszi távolság 202,538).

d) Az a) pont beállításait annyiban módosítjuk a klaszterfuttatás előtt, hogy a 'Method...' menüpont 'Cluster Method' dobozában a 'Between-groups linkage' opciót választjuk.
A csoportok közti átlagos lánc módszerrel megváltozik az ügyfelek, illetve az általuk alkotott klaszterek összevonása. Ez a módszer a csoportok közötti távolságot úgy határozza meg, hogy veszi az adott két csoport valamennyi elemének távolságát (például, ha az egyik csoportban 3, a másik csoportban 2 elem van, akkor összesen 6 távolságértéket határoz meg), majd azokat átlagolja. Azt a két klasztert vonja össze, melyek között a legkisebb a távolság.

A dendrogramon 40%-os távolságszintnél csak két klaszter látható, a 2. klaszter továbbra is a kiugró 445-ös ügyfelet tartalmazza.
Érdemes újra lefuttatni a két eljárással a hierarchikus klaszterelemzést úgy, hogy a 'Save' menüpontban a 'Single Solution' lehetőséget választva megadjuk a 40%-os távolságszint mellett kapott klaszterszámokat. Ezzel az adattáblába új változóként elmenthetjük az egyes ügyfelek klaszterbesorolásait a 40%-os távolságszinten létrejött klaszteknek megfelelően. A két futtatás eredményeként kapott új változókat - mivel nominális változók - az **Analyze → Descriptive Statistics → Crosstabs...** parancson belül asszociációs mérőszámokkal célszerű vizsgálni, hogy a felosztások hasonlóságát megállapíthassuk.
Az egyes ügyfelek közötti távolságok nem változnak, hiszen a klaszterezés, azaz az egyes klaszterek összevonásának módszerét változtattuk meg, nem az egyes ügyfelek közötti távolságszámítási eljárást. Így a legközelebbi és legtávolabbi ügyfelek és a közöttük lévő távolság mértéke megegyezik a feladat c) pontjában adott válasszal.

4. feladat:

a) Ismételjük meg a 3. feladat a) pontját úgy, hogy az esetek helyett a változókat klaszterezzük! Hány csoportba lehetséges tömöríteni a változókat?

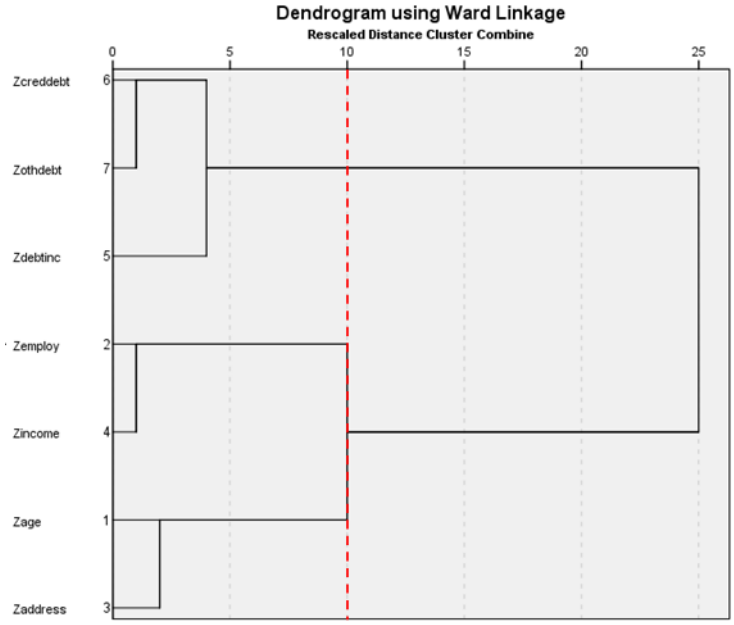
b) Elemezze a dendrogramot! Milyen sorrendben vonja össze a változókat a klaszterelemzés? Mely változókat lenne célszerű összevonni?

c) Melyik két változó van legközelebb egymáshoz? Melyik kettő van legtávolabb egymástól? Mekkorák a távolságok?

A feladat megoldása:

a) A 3. feladat a) pontjának beállításait annyiban módosítjuk a klaszterfuttatás előtt, hogy a 'Cluster' dobozban a 'Variables' opciót választjuk a hierarchikus klaszterezés menüponton belül.
A bevont hét változó kettő vagy három homogén csoportot alkot, mivel a dendrogram 40%-os távolságszintjénél (a 25 egységre sztenderdizált távolság 10-es egységénél) a 2. és a 3. csoport összevonódik.

b) A dendrogramon látható, hogy az első lépésekben a 'creddebt' és az 'othdebt', valamint az 'employ' és az 'income' változók kapcsolódnak össze. Ezt követően az 'age' és az 'address' változót vonja össze az eljárás egy klaszterbe, majd a 'debtinc' változó hozzákapcsolódik a 'creddebt' és az 'othdebt' változók csoportjához. A távolságszint 40%-ánál kapcsolódik össze az 'employ' és az 'income', valamint az 'age' és az 'address' változók csoportja. Az így létrejött két klaszter jól elkülönül egymástól.
A dendrogram alapján a bevont hét változót három csoportba lehetne összevonni: adóssággal ('creddebt', 'othdebt', 'debtinc'), foglalkoztatással ('employ', 'income'), személyes jellemzőkkel ('age', 'address') kapcsolatos klaszterek.



c) A távolságmátrixból ('Proximity Matrix') kiolvashatjuk, hogy a legközelebbi változók a 'creddebt' és 'othdebt' változók (köztük lévő négyzetes euklideszi távolság 602,871), a legtávolabbi változók 'debtinc' és 'address' (köztük lévő négyzetes euklideszi távolság 1753,930).

Gyakorló feladatok

1. Végezzen nemhierarchikus klaszterelemzést a fizetőképes ügyfelekre ('default' változó = 0) a 2. feladat a) részében megjelölt sztenderdizált változók bevonásával! A 'könyökszabály' segítségével állapítsa meg hány klasztert érdemes kérni! Az optimális csoportszámmal kapott eredményeket elemezze a 2. feladat b), c) és d) pontja alapján!
2. Ismételje meg a 3. feladat a) pontjának klaszterelemzését azzal a módosítással, hogy a négyzetes euklideszi távolság helyett más távolságmérést alkalmaz! Melyik két főiskolai végzettségű ügyfél van legközelebb egymáshoz? Melyik két ügyfél van legtávolabb egymástól? Mekkorák a távolságok? Több távolságmérési lehetőséget is próbáljon ki! Hasonlítsa össze az egyes távolságmérési eljárások eredményeit! Vizsgálja meg a kapott dendrogramokat!
3. Válaszolja meg a 3. feladat d) pontját a 'between-groups linkage' módszer helyett más módszert alkalmazva! Valamennyi SPSS által felkínált klaszterezési módszert próbálja ki! Az egyes módszereknél hogyan változik a dendrogram struktúrája? Hány csoportra lenne érdemes felbontani a főiskolai végzettségű ügyfeleket az egyes módszerek eredménye alapján?

Irodalomjegyzék

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. Melyek a hierarchikus és nemhierarchikus klaszterelemzés különbségei?

   a) Az alkalmazás célja: az egyiket inkább feltáró, a másikat inkább megerősítő elemzésként használják.
   b) A hierarchikus klaszterelemzés során konkrét elképzelésünk van a csoportok számára (k nagyságára) vonatkozóan, míg a nemhierarchikus klaszterezésnél ennek lehetséges nagyságának megállapítása a cél.
   c) A nemhierarchikus klaszterelemzés két megközelítésben végezhető, gyakorlati alkalmazásokban az összevonó eljárás az elterjedtebb.
   d) Mindegyik előző válasz helyes.

2. Mit jelent a stabilitás követelménye a klaszterezés esetén?

   a) A bevont változók relatív szórása ne haladja meg a kettőt.
   b) Az adatokban bekövetkező kis változások kis változást eredményezzenek a felosztásban.
   c) Ha egy egyedet elveszünk vagy hozzáadunk a megfigyelésekhez, akkor az osztályozásban nagyon kis változás következzen be.
   d) Egyik előző válasz sem helyes.

3. Mely állítás igaz a klaszterelemzéssel kapcsolatban?

   a) A dendrogram olyan speciális két- vagy többdimenziós ábra, mely megmutatja, hogy az egyes elemek milyen távolságmérték mellett kapcsolódnak össze.
   b) A hieararchikus klaszterelemzésnél alkalmazható Ward módszer egyenlő elemszámú klaszterek kialakítására törekszik.
   c) A k-középpontú klaszterezés során a lehetséges klaszterek maximum számát a $\sqrt{n/2}$ képlet segítségével határozhatjuk meg.
   d) Pontosan két állítás igaz.

4. Melyik állítás igaz a hierarchikus klaszterelemzésnél alkalmazható egyszerű lánc módszerre?

   a) Tértágító hatású
   b) Alkalmazásakor inverzió léphet fel
   c) Jellemzője a lánchatás, azaz az elemek csak közvetlenül, láncszerűen kapcsolódhatnak össze.
   d) Egyik előző válasz sem helyes.

5. Milyen célt szolgál az ANOVA tábla (szóráselemzés) a nem-hierarchikus klaszterelemzésben?

   a) Segítségével kiválaszthatjuk a csoportokat elkülönítő változókat.
   b) Megmutatja, hogy a változók együttesen szignifikáns szerepet játszanak-e az osztályozásban.
   c) Mindkét előző válasz helyes.
   d) Egyik előző válasz sem helyes.

4. fejezet

LINEÁRIS REGRESSZIÓ

A módszer rövid összefoglalása

Talán a legismertebb többváltozós elemzési módszer, amelynek során egy kiválasztott arány skálán mért eredményváltozó értékét közelítjük magyarázó változók lineáris kombinációjával. Az együtthatók becslése a legkisebb négyzetek módszerén alapul. Amennyiben nagyszámú változó áll rendelkezésünkre, az elemzésünket az SPSS-ben jelentősen megkönnyíti a stepwise módszertan, amelynek révén a lehető legpontosabb becslést eredményező magyarázó változó kombináció is kiválasztásra kerül a megadott változók köréből (vagyis a modellezés során nem szükséges az összes lehetséges regressziót becsülnünk és ezáltal az elvárt tulajdonságokat figyelembe véve meghatározni közülük a legmegfelelőbbet).

Megoldási módszerek és az eredmények értelmezése

A gyakorló feladatok megoldásánál említett változók a car_sales.sav adatai között találhatók.

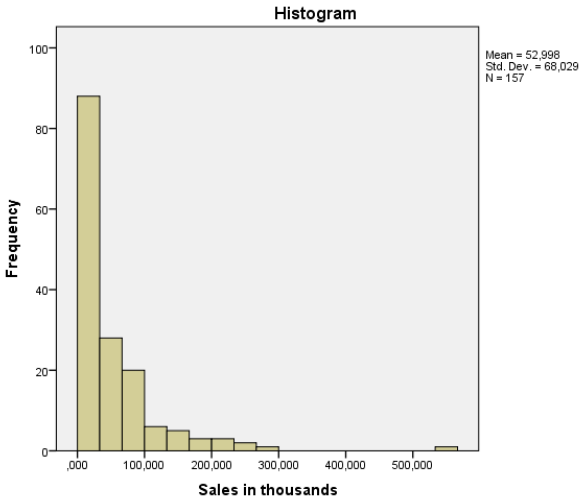
1. feladat:
Futtasson Enter típusú regressziót az értékesítés (sales) változóra az alábbi magyarázó változók mellett:

a, price;
b, price, horsepower, engine size
c, futassa le a b, feladatot Stepwise módszerrel.

Értelmezze a kapott eredményeket, amennyiben lehetséges: modell magyarázóereje, együtthatók, multikollinearitás, kilógó értékek, reziduálisok. A futtatás előtt, amennyiben előnyös, transzformálja a magyarázott változót. Ha igen, akkor miért? Milyen transzformációt alkalmazna?

A feladat megoldása:

Az elemzés elején vizsgáljuk meg a sales változó normalitását. Mind a Kolmogorov-Smirnov, mind a Shapiro-Wilk teszt alapján elvetjük a normalitást. A változó hisztogramja egy erősen csúcsos és jobbra hosszan elnyúló eloszlást mutat:



Tests of Normality

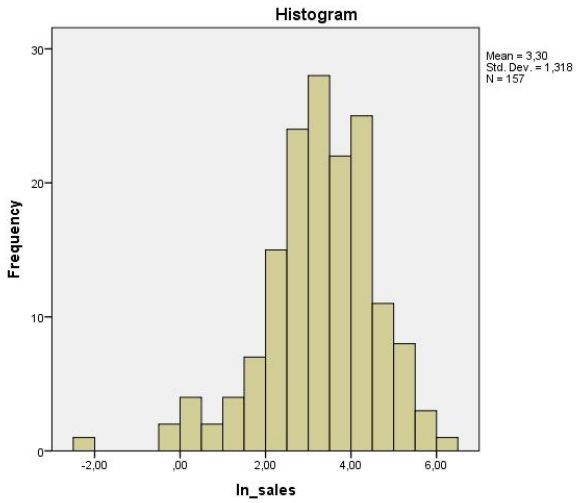
	a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Sales in thousands	,218	157	,000	,667	157	,000

Lilliefors Significance Correction

Bár a lineáris regresszió legkisebb négyzetekkel való becslésének nem feltétele a magyarázott változó normalitása, sok alkalmazott teszt jobb illeszkedést mutat, ha teljesül a normalitás. Erősen jobbra ferde változók esetén a logaritmizálás segíthet a normalizálásban. A változó természetes alapú logaritmusát ln_sales néven a következő menüpont választásával lehet számolni:

Transforme → Compute Variable

Ekkor a transzformált változó hisztogramja és a Kolmogorov-Smirnov teszt alapján (5%-os szignifikancia szinten) normális eloszlásúnak tekinthető:



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ln_sales	,066	157	,093	,964	157	,000

Lilliefors Significance Correction

Az **a, kérdés** regressziójának becslésénél, mivel egy magyarázó változónk van, számos tulajdonságra nem kell figyelnünk (pl. multikollinearitás). Ekkor Enter módszerrel a következő eredmények adódnak:

A becslés során alkalmazott változók leíró statisztikái nem mutatnak mintán belüli szeparációt, a relatív szórások egyik esetben sem érik el az egyet. Ez különösen fontos, hiszen amennyiben valamelyik változó mentén a minta szeparálódna, adott esetben a mintát kettévágva két különböző regressziós becsléssel pontosabb eredményhez juthatnánk. A kapott eredmények alapján az adataink koncentráltnak tekinthetők, mivel a relatív szórás (szórás/átlag) mind az eredmény mind a magyarázó változó esetén alacsony értéket vesz fel. A 2. táblázat alapján az ln_sale változó relatív szórása 0,4, míg a price változóé 0,52.

Descriptive Statistics

	Mean	Std. Deviation	N
ln_sales	3,2957	1,32457	155
Price in thousands	27,39075	14,351653	155

A kapott regresszió R^2 értéke 0,305, a korrigált R^2 is mindössze 0,301, ami nem jelez túl erős determináltságot. Az ANOVA tábla F-tesztje alapján a price változó szignifikáns a becslés szempontjából. A becsült regressziós egyenlet a következő:

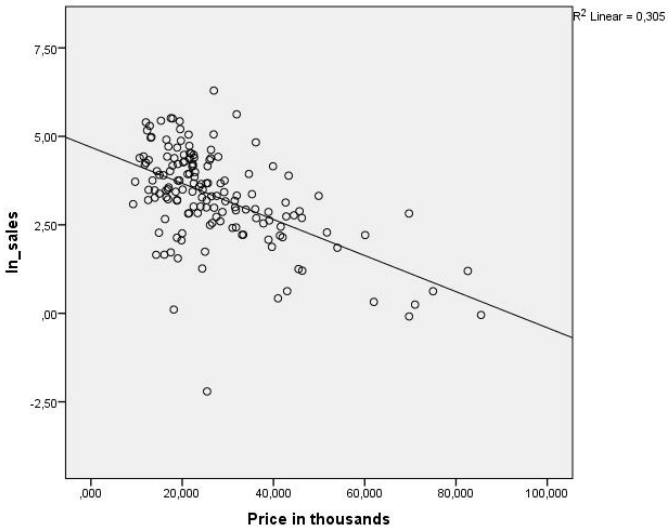
$$\ln\_sales = -0{,}051*price + 4{,}692$$

ami átalakítva

$$sales = e^{-0{,}051*price} * e^{4{,}692}$$

vagyis egy adott autótípus árának 1000 dolláros növekedése $(1/e^{0{,}051}) = 0{,}95$-szeresére csökkenti az eladott mennyiséget. Az adatbázisunk tehát visszaigazolja azt az általánosnak is tekinthető megállapítást, hogy a drágább autókból kevesebbet vásárolnak. (A reziduálisok és a kilógó pontok elemzésére a b, pont megoldásánál térünk ki.)

A kapott eredményeket egy kétdimenziós ábrán szemléltetve:



A **b, pontnál** a lineáris regresszió Enter módszerrel való becsléséhez további magyarázó változóként vonjuk be a horsepower és az engine size változókat. A kapott regresszió R^2 értéke 0,382, a korrigált R^2 0,37, mely továbbra is alacsony determináltságot jelez. A további két magyarázó változó bevonásával a modell magyarázó ereje alig módosult, ami multikollinearitásra is utalhat.

b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	a	,382	,370	1,05139

Predictors: (Constant), Engine size, Price in thousands, Horsepower

Dependent Variable: ln_sales

a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4,127	,318		12,990	,000		
	Price in thousands	-,061	,011	-,659	-5,402	,000	,275	3,637
	Horsepower	-,005	,004	-,209	-1,198	,233	,135	7,419
	Engine size	,568	,153	,449	3,700	,000	,278	3,603

Dependent Variable: ln_sales

A becsült regressziós egyenlet:

$$\ln\_sales = 4{,}127 - 0{,}061 * price - 0{,}005 * horsepower + 0{,}568 * engine\ size.$$

A t-statisztika szignifikanciaszintjéből látható, hogy a horsepower változó nem szignifikáns, nem vethető el a $\beta=0$ nullhipotézis.

A price és az engine size esetén a VIF értékek meghaladják a 3-at, amely már zavaróan nagy multikollinearitásra utal. A horsepower 7,419-es VIF szintje (5 feletti)

pedig már használhatatlan modellt jelez. A kondíciós indexek 21,491-es értéke (15 feletti) szintén az elfogadhatatlan mértékű multikollinearitást mutat.

A multikollinearitási statisztikák alapján használhatatlan modellt kaptunk. Az Enter módszer alkalmazásánál erre a becslés során feltétlen figyelni kell. Ennek kiküszöbölésére alkalmazható lehet a Stepwise módszertan, ahol a magyarázó változók megadott feltételek alapján kerülnek be- illetve kiléptetésre (de a Stepwise módszer alkalmazásakor is előforduhat jelentős mértékű multikollinearitás egy lineáris regressziós modellben). Az Enter és a Stepwise közötti választást elsősorban az elemzés célja határozza meg:

- megerősítő elemzéseknél az Enter;
- feltáró elemzéseknél a Stepwise alkalmazandó.

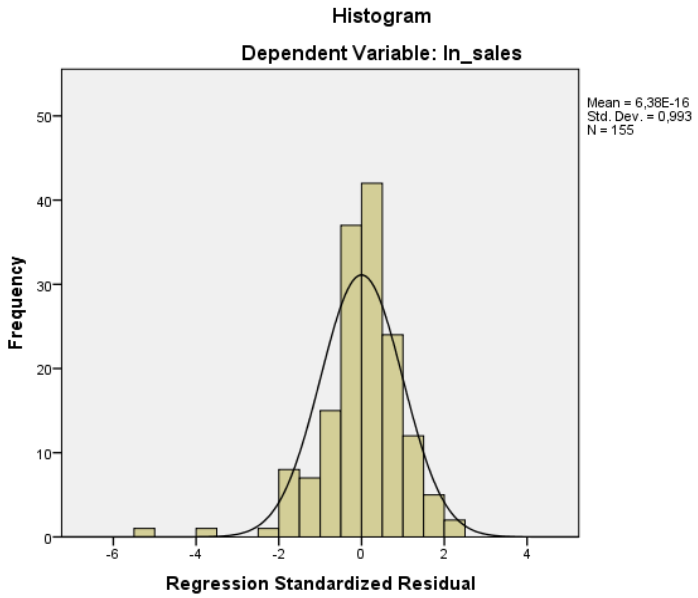
A **c, kérdés**ben Stepwise módszerrel lefuttatva két magyarázó változó kerül bevonásra a regressziós egyenletbe.

a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4,692	,192		24,417	,000		
	Price in thousands	-,051	,006	-,553	-8,200	,000	1,000	1,000
2	(Constant)	3,911	,262		14,934	,000		
	Price in thousands	-,071	,008	-,767	-9,330	,000	,607	1,647
	Engine size	,432	,104	,342	4,162	,000	,607	1,647

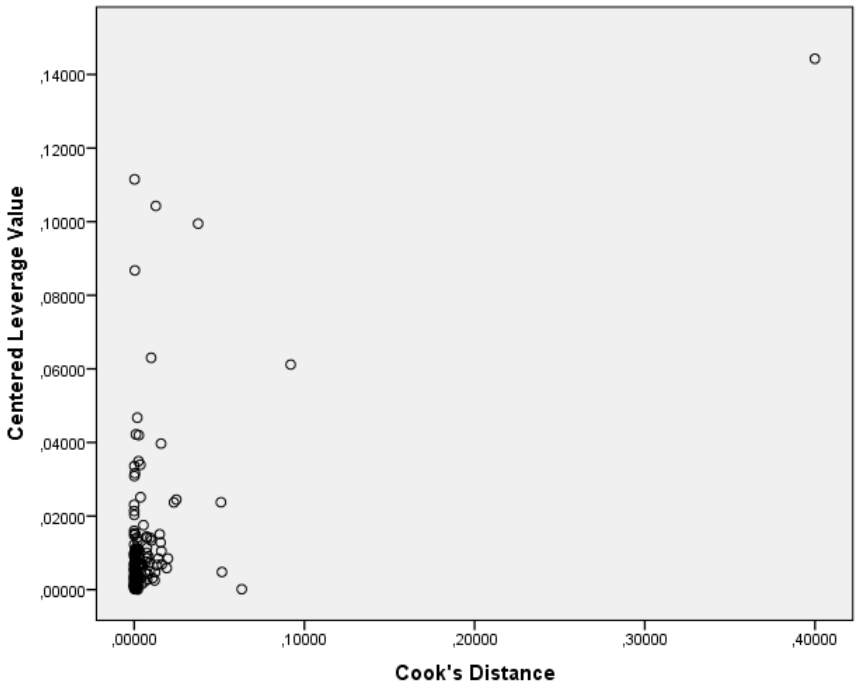
Dependent Variable: ln_sales

A t-statisztikák alapján mindkét bevont magyarázó változó szignifikáns. Az együtthatójuk ellentétes előjelű. Az ár (price) növekedése csökkenti, a motorméret (engine size) növekedése pedig növeli az eladott mennyiséget. A kollinearitási statisztikák elfogadható modellt jeleznek: VIF<2, tolerancia>0,5.



A sztenderdizált reziduálisok ábrája a normális eloszlásnál csúcsosabb, de csak negatív irányban utal kilógó értékekre. A Cook-távolság és a Leverage értékeket

megvizsgálva azonban nem tudunk kilógó pontokat azonosítani, mivel egy elem esetén sem kerül elérésre a kritikusnak tekintett 1 (Cook) és 0,2-es (Leverage) szint.



Gyakorló feladatok

1. Végezzen Stepwise lineáris regressziós elemzést a bankloan.sav fájl alábbi változóira: income - eredményváltozó, address, age, creddebt – magyarázó változók!

2. Mennyiben változik a lineáris regressziós becslés eredménye, ha stepwise módszertannál az eredeti változók helyett sztenderdizált változókból indulunk ki?

3. Mennyiben változik a lineáris regressziós becslés eredménye, ha stepwise módszertannál az eredeti változók helyett centrált változókból indulunk ki?

4. Tegyük fel, hogy egy lineáris regressziós elemzésben a Cook távolság értékek alapján egyetlen kilógó érték jelenlétére lehet következtetni. Mire következtet ebből az elemzés további menetére vonatkozóan?

5. Egy feladatban a német bankszektorra készít elemzést lineáris regresszió segítségével. A regressziós becslést egy kilógó pont jelentősen befolyásolja. Elhagyná-e a kilógó bankot a további elemzésből? Segítségül gondolja meg, mi lenne a válasza, amennyiben a kilógó bank a Deutsche Bank vagy egy kis tartományi pénzintézet.

6. Az előző feladathoz hasonlóan most is a német bankrendszert elemzi. Változónak felhasználja a bankok kamateredményeit, amely alapján több kilógó pont keletkezik, és megkérdőjelezhető a regresszió során a változó felhasználása. Milyen módon tartható meg a kamateredmény változóból származó információ az elemzésben? Soroljon fel néhány lehetséges megoldást!

Irodalomjegyzék

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. A (nem korrigált) R-négyzet maximális értéke

   a) az elemzésbe bevont változók számával egyenlő
   b) egy
   c) nulla
   d) egyik előző válasz sem helyes.

2. A lineáris regressziós modell egyik alkalmazási előfeltevése, hogy

   a) az elemzésbe bevont változók szórása megegyezik
   b) az elemzésbe bevont változók átlaga megegyezik
   c) az elemzésbe bevont változók kovarianciamátrixa diagonális
   d) egyik előző válasz sem helyes.

3. Egy lineáris regressziós modellben a VIF értékek

   a) maximális értéke egy
   b) minimális értéke nulla
   c) mindkét előző válasz helyes
   d) egyik előző válasz sem helyes.

4. Egy modellben a kondíciós indexek

   a) száma megegyezik az elemzésbe bevont (magyarázó) változók számával
   b) összege megegyezik az elemzésbe bevont (magyarázó) változók számával
   c) mindkét előző válasz helyes
   d) egyik előző válasz sem helyes.

5. fejezet

LOGISZTIKUS REGRESSZIÓ

A módszer rövid összefoglalása

Az általánosított lineáris modellek (GLM - General Linear Models) közé tartozó logisztikus regresszió fontos jellemzője, hogy a függő változó nem folytonos, hanem diszkrét változó. Osztályozó (klasszifikációs) eljárások közé sorolható, mivel akkor alkalmazzuk, ha előre definiált, egymást kölcsönösen kizáró csoportok egyikébe soroljuk be a megfigyeléseket a magyarázó változókból nyert információ alapján. Kétfajta logisztikus regressziót használhatunk: bináris (a megfigyelt eseménynek csak két állapota van) vagy polichotom (a megfigyelt esemény több állapotú) regressziót. A továbbiakban csak a bináris regresszióval foglalkozunk.

A bináris regresszió azt tételezi fel, hogy a magyarázó változók az egyik kimenetel (pl. a csőd, fizetőképtelenség, kárbekövetkezés stb.) bekövetkezési esélyét magyarázzák. Az Y dichotom változó, ahol Y=1 az esemény bekövetkezését jelöl (pl. fizetőképtelenség), p pedig ennek a bekövetkezési valószínűségét [p(Y=1)]. Az esély vagy odds értékét az alábbi képlettel határozzuk meg, ami az X-től (magyarázó változóktól) függő feltételes valószínűségek aránya.

$$odds = \left(\frac{p}{1-p} \right) = \exp(b_0 + b_1 x_1 + ... + b_p x_p) = e^{\underline{bx}}$$

A magyarázó változók között lehetnek nominális, ordinális vagy magasabb, intervallum és arányskálán mért változók is.
Az esély logaritmusa a logit, ami a magyarázó változók lineáris függvénye:

$$\ln(odds) = \ln\left(\frac{p}{1-p} \right) = \log it(p) = b_0 + b_1 x_1 + ... + b_p x_p$$

A p valószínűség 0 és 1 közötti tartományban mozoghat, legbizonytalanabb értéke 0,5. Az esély vagy odds 0 és végtelen között, itt az 1 a bizonytalanságot jelző érték, és a tartomány nem szimmetrikus. Az odds logaritmusa (logit)]-∞; +∞[között veheti fel értékeit, a 0 érték jelenti a bizonytalanságot.

A magyarázó változók együtthatóinak értelmezésénél a b_i becsült paraméter az x_i változó egy egységnyi abszolút, ceteris paribus változásának a logitra gyakorolt parciális hatását mutatja, közvetlen tartalma nincs. Az $\exp(b_i)$ az x_i egy egységnyi abszolút növekedésének ceteris paribus hatása az odds-ra, hányszor nagyobb az „Y=1" bekövetkezésének esélye.

Az Y eredményváltozó kategóriáinak bekövetkezési valószínűsége [p(Y=1)] az X magyarázó változókból nem becsülhető a hagyományos legkisebb négyzetek módszerével, ehelyett a Maximum Likelihood (ML) becslést alkalmazzuk.

Megoldási módszerek és az eredmények értelmezése

A gyakorló feladatokat a bankloan.sav adatain végezzük el.

1. feladat:
Bináris logisztikus regresszióval szeretnénk vizsgálni és magyarázni, hogy a banki ügyfelek körében a fizetőképtelenséget ('Prevously defaulted [default]' változó) mely tényezők milyen mértékben befolyásolják.

Végezzen logisztikus regressziós elemzést (forward Wald módszer) az 'age', 'ed' (kategóriaváltozó), 'employ', 'address', 'income', 'debtinc', 'creddebt', 'othdebt' változók alapján a 'default' változó két kategóriája esetében (cut value: 0,5).
Az eredményváltozónak melyik kategóriája a kontroll csoport és melyik kategóriáját magyarázzuk a modellel?

A feladat megoldása:

A logisztikus regresszió a következő menüpont kiválasztásával érhető el:

Analyze → Regression → Binary logistic...

A 'Dependent' dobozba áthelyezzük az eredményváltozót, ami csak és kizárólag egy két érték-kategóriával rendelkező változó lehet, jelen esetben a 'Previously defaulted [default]'. A 'Covariates' dobozba áthelyezzük a megadott magyarázó változókat. A 'Method'-ban tudjuk beállítani a kért beléptetési módot, alapbeállítása az 'Enter' beléptetés.
A jobboldali 'Categorical...' menüpontban a baloldali 'Covariates' dobozból helyezzük át a jobboldali 'Categorical Covariates' dobozba a nominális és ordinális skálájú változókat, jelen esetben csak a 'Level of education [ed]' változót[5], majd 'Continue' gombbal visszajutunk az előző ablakba.

A 'Dependent Variable Encoding' táblázatból tudjuk, hogy az eredményváltozó 1 értéke a 'Yes' kategóriának feleltethető meg, tehát a modell a fizetőképtelen ügyfelek ('default' = 1) csoportjába kerülést magyarázza, a kontroll csoportban pedig a fizetőképes ügyfelek vannak.

2. feladat:
A 'Level of education [ed]' változó egyes kategóriái az 1. feladatban illesztett modell mely változóinak feleltethetők meg?

A feladat megoldása:

A 'Categorical Variables Coding' táblázat mutatja az 'ed' változó egyes kategóriáinak gyakoriságát, illetve a hozzájuk tartozó kategóriakódolást (oszlopokban a modellbe

[5] A 'Change Contrast' dobozban beállíthatjuk, hogy az adott kategóriás változót hogyan kezelje (itt az 'Indicator' beállítás általában megfelelő') az SPSS, illetve hogy a változó első vagy utolsó kategóriáját tekintse referencia kategóriaként, amihez a többi kategóriát hasonlítja majd a modell. Ez az eredmények értelmezésénél lesz fontos. Célszerű előzetesen megvizsgálni a változó egyes kategóriáiban a megfigyelési egységek számát (pl. Explore paranccsal), és azt választani, amely kategóriában van elegendő esetszám. Az első és az utolsó kategóriát ABC sorrend alapján azonosítja az SPSS. Szükség esetén a kategóriák átkódolásával állíthatjuk be a kívánt referencia kategóriát. A 'Change' gombbal hagyjuk jóvá módosításainkat.

kerülő változók, amiket a további kimenetekben ed(1), ed(2), ed(3), ed(4) névvel azonosít az SPSS, sorokban az 'ed' változó egyes kategóriái). Az 'ed' kovariánsnak öt kategóriája van, amit elegendő négy (0-1 értékű) változóval leírni (különben egzakt multikollinearitás lépne fel). A modell futtatási beállítása során az utolsó kategóriát állítottuk be referencia kategóriának, így a 'Post-undergraduate degree' kategória paraméter kódolása csak nullákat tartalmaz. Az 'ed1' változó kódolásánál a 'Did not complete high school' kategória vesz fel 1 értéket, a többi kategória értéke nulla. Ez alapján a későbbiekben az 'ed(1)' változó mint magyarázó változó azt mutatja meg, hogy a 'Did not complete high school' kategóriába tartozók a 'Post-undergraduate degree' kategóriába tartozókhoz képest hogyan módosítják az eredményváltozót. Az 'ed2', 'ed3', 'ed4' változók értelmezése ehhez hasonlóan történik.

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Level of education	Did not complete high school	372	1,000	,000	,000	,000
	High school degree	198	,000	1,000	,000	,000
	Some college	87	,000	,000	1,000	,000
	College degree	38	,000	,000	,000	1,000
	Post-undergraduate degree	5	,000	,000	,000	,000

3. feladat:

Értékelje az 1. feladatban kapott modell együtthatóit! Mely változók szignifikánsak? Mi a szignifikáns együtthatók tartalma? Melyik szignifikáns változó növeli, illetve csökkenti leginkább a fizetőképtelenség esélyét? Értelmezze ezen változóknak az exp(b) együtthatóit és a konfindencia intervallumukat!

A feladat megoldása:

Egy együttható szignifikanciáját a Wald teszt p-értéke alapján döntjük el, ahol a H_0 azt jelenti, hogy az együttható 0 (azaz nem szignifikáns a változó).

A Block 1: Method = Forward Stepwise (Wald) szakasz 'Variables in Equation' táblázata alapján értékeljük a becsült együtthatókat. Mivel változószelekciós eljárást alkalmaztunk, a táblázat megmutatja lépésről lépésre a bevont változókat. A választott eljárás által véglegesnek tekintett modellt a 4. lépésben kapjuk meg, amely a – forward stepwise eljárás beállításainak megfelelően – csak az 5%-on szignifikáns változókat tartalmazza: az 'employ', az 'address', a 'debtinc' és a 'creddebt' változókat, illetve a konstanst. Az 'employ' és az 'address' változók együtthatója negatív előjelű, ezen változók értékének egységnyi növekedése ceteris paribus csökkenti a fizetőképtelenség bekövetkezésének (azaz default változó értéke = 1) esélyét. A többi szignifikáns változó pozitív előjele arra utal, hogy azok értékének egységnyi növekedése ceteris paribus növeli a fizetőképtelenség bekövetkezésének esélyét.

A fizetőképtelenség bekövetkezésének esélyére a legerősebb pozitív hatással a 'creddebt' változó (Exp(b) = 1,774, $p < 0,001$), a legerősebb negatív hatással az 'employ' változó (Exp(b) = 0,785, $p < 0,001$) van.

A 'creddebt' változó exp(b) együtthatója azt mutatja meg, hogy ha a hitelkártya-adósság 1 000 dollárral növekszik, ceteris paribus 1,774-szorosára (vagy 77,4%-kal) növeli a fizetőképtelenség bekövetkezésének esélyét. A változóhoz tartozó 95%-os konfidencia intervallum alapján elmondható, hogy 95%-os megbízhatósági szint

mellett a 'creddebt' változó tényleges hatása az 1,495 és 2,104 intervallumba esik - távol az 1 értéktől, ami a változó semleges hatását jelezné.
Az 'employ' változó exp(b) értéke alapján megállapíthatjuk, hogy a jelenlegi munkahelyen töltött munkaévek számának növekedése ceteris paribus 0,785-szörösére változtatja (vagy 21,5%-kal csökkenti) a fizetőképtelenség bekövetkezésének esélyét. Az együtthatóhoz tartozó 95%-os konfidencia intervallum alsó és felső hatása 0,743, illetve 0,829.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 4[a]	employ	-,243	,028	74,761	1	,000	,785	,743	,829
	address	-,081	,020	17,183	1	,000	,922	,887	,958
	debtinc	,088	,019	22,659	1	,000	1,092	1,053	1,133
	creddebt	,573	,087	43,109	1	,000	1,774	1,495	2,104
	Constant	-,791	,252	9,890	1	,002	,453		

a. Variable(s) entered on step 4: address.

4. feladat:

Értékelje az 1. feladatban kapott modell osztályozásának jóságát! Hogyan tudnánk ezt javítani anélkül, hogy a modellt módosítanánk?

A feladat megoldása:

A modell osztályozásának jóságát a klasszifikációs táblázat ('Classification Table') segítségével értékelhetjük. Ez tulajdonképpen egy speciális kereszttábla, ami az eredményváltozó (jelen esetben a fizetőképtelenség szerinti besorolás) tényleges és a becsült értékeinek együttes eloszlását mutatja.

A csak konstanst tartalmazó modell teljes találati aránya[6] (lásd 'Block 0: Beginning Block' 'Classification Table' részen) 73,9%. Az illesztett, változókat is tartalmazó modell teljes találati aránya 81,4%-ra javul. A modell a 'Previously defaulted' No kategóriájába tartozó (azaz fizetőképes) ügyfeleket 92,5%-ban sorolja be helyesen, a Yes kategóriába tartozókat (azaz fizetőképteleneket) már csak 50,3%-ban. A fizetőképtelen ügyfelek közel felének téves besorolása alapján nem lehetünk meggyőződve arról, hogy jó modellt kaptunk. (Gondoljunk csak arra, hogy mi történne akkor, ha valamely bank erre a modellre alapozva döntene arról, hogy mely ügyfeleknek ad újabb hiteleket.) Kérdés, hogy vajon ezt lehet-e javítani a modell módosítása nélkül vagy csak annak módosításával.

Classification Table[a]

			Predicted		
			Previously defaulted		Percentage Correct
	Observed		No	Yes	
Step 4	Previously defaulted	No	478	39	92,5
		Yes	91	92	50,3
	Overall Percentage				81,4

a. The cut value is ,500

Az osztályozás jósága változtatható a modell módosítása nélkül az ún. vágási érték (’cut value’) változtatásával. Eddig a modell futtatásánál 0,5 cut value-t alkalmaztunk.

[6] A teljes találati arány azt mutatja meg, hogy az összes megfigyelés hány százalékát sikerült helyesen besorolnia a modellnek.

Ez azt jelenti, hogy a modell az egyes megfigyelésekhez (itt ügyfelekhez) becsült valószínűségi értékek ('Predicted probability [PRE_1]' változó[7]) alapján besorolja a fizetőképes-fizetőképtelen csoportokba az ügyfeleket, mégpedig úgy, hogy ha a PRE_1 változó < 0,50, akkor a 0-as (No, azaz fizetőképes) csoportba kerül az ügyfél, ellenkező esetben az 1-es (Yes, azaz fizetőképtelen) csoportba. Így kapjuk meg a 'Predicted group [PGR_1]' változót[8], amit a klasszifikációs táblában összevet az eredeti csoportbesorolással (a Previously defaulted [default]' változóval).
Célszerű több, különböző vágási értékkel ismételten futtatni a modellt[9]. Ezek kiválasztásában az output 'Observed Groups and Predicted Probabilities' ábrája lehet segítségünkre. A vágási érték csökkentésével csökken a fizetőképes ügyfelek találati aránya, a fizetőképteleneké viszont nő. Például 0,4-es vágási értéknél a fizetőképes ügyfelek találati aránya 86,5%-ra csökken, a fizetőképteleneké viszont 62,3%-ra nő, összességében azonban nem javul a találati arány (80,1%). Próbáljon ki más vágási érték beállításokat is, és hasonlítsa össze az eredményeket a példa alapján!
Ahogy tapasztalhattuk, a klasszifikációs tábla vágási értéktől való függése befolyásolja az osztályozása jóságának értékelését. Ennek kiküszöbölésére használható pl. az ROC (Receiver Operating Curve) görbe, ami minden lehetséges vágási értéket figyelembe vesz (ROC görbéről lásd 7. feladat) és annak segítségével vizsgálható a klasszifikáció jósága.

5. feladat:

Értékelje az 1. feladatban kapott modell illeszkedését! Hogyan lehet mérni a modell jóságát?

A feladat megoldása:

Az illesztett modell alkalmazhatóságáról ad információt a Hosmer-Lemeshow teszt, amely a megfigyeléseket a becsült valószínűségek alapján g számú csoportra, általában decilisekre (g=10) osztja. Azt vizsgáljuk, hogy a decilisekre a ténylegesen bekövetkező (megfigyelt - M) események száma megegyezik-e az előrejelzettel (várt - V) a bináris változó kategóriáiban. A homogenitásvizsgálat tesztstatisztikája az alábbi módon írható fel, amely (g-2) szabadságfokú khi-négyzet eloszlást követ:

$$\chi^2 = \sum (M - V)^2 / \left(V(1 - \sum p / s\right)$$

Ha szignifikáns eltérés van, akkor nem jó a modell illeszkedése.

A teszt elvégzéséhez futtassuk újra a modellt az 'Options...' menüpontban a 'Hosmer-Lemeshow goodness-of-fit' lehetőséget kérve. Az eredmények alapján megállapíthatjuk, hogy a teszt nullhipotézisét nem tudjuk elutasítani (p érték 0,381) a szokásos szignifikancia szintek mellett, a modell illeszkedésében nem találtunk szignifikáns eltérést az előrejelzett értékektől, a modell illeszkedése elfogadható.

[7] Az egyes megfigyelésekhez tartozó becsült valószínűségek mentését a logisztikus regresszió beállításainál a 'Save' parancson belül a 'Probablities' lehetőség választásával kérhetjük. Ismételt futtatás után az adattábla megjelenik a PRE_1 változó.

[8] Az egyes megfigyelésekhez tartozó, modell által becsült csoportbasorolás mentését a logisztikus regresszió beállításainál a 'Save' parancson belül a 'Group membership' lehetőség választásával kérhetjük. Ismételt futtatás után az adattábla megjelenik a PGR_1 változó.

[9] Az ismételt futtatás során érdemes a mentési opciókat ('Save...' fülre kattintva) átállítani és nem menteni ismét a korábban kért, nem módosuló változókat.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
4	8,556	8	,381

Contingency Table for Hosmer and Lemeshow Test

		Previously defaulted = No		Previously defaulted = Yes		
		Observed	Expected	Observed	Expected	Total
Step 4	1	70	69,669	0	,331	70
	2	69	68,554	1	1,446	70
	3	64	66,539	6	3,461	70
	4	64	63,521	6	6,479	70
	5	65	59,692	5	10,308	70
	6	50	55,141	20	14,859	70
	7	48	49,016	22	20,984	70
	8	43	41,000	27	29,000	70
	9	32	30,470	38	39,530	70
	10	12	13,397	58	56,603	70

A modell jóságát két R-négyzet jellegű mutató - Cox & Snell R Square és a Nagelkerke R Square - méri, amiket a 'Model Summary' táblázatból olvashatunk ki. Az R^2 már nem értelmezhető a lineáris regressziónál megszokott módon, a megmagyarázott variancia százalékaként, csupán annyit mond, hogy a csak konstanst tartalmazó (null)modellhez tartozó log likelihood értéket hány százalékkal sikerült csökkenteni. Mindkét mutató értéke 0 és 1 közé esik alapesetben. Minél nagyobb a mutatók értéke, annál jobb a modell illeszkedése. A kapott értékek alapján a modell jósága elfogadható.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
4	556,732[a]	,298	,436

6. feladat:

Az 1. feladat modellje révén becsült valószínűségek közül melyik a legkisebb és melyik a legnagyobb?

A feladat megoldása:

A feladat megválaszolásához futtassuk újra a modellt a 'Save...' menüpontban a 'Probalities' lehetőséget kérve.

Az adattáblába mentett becsült valószínűségekre (PRE_1 nevű új változó) futtathatjuk valamelyik leíró statisztika parancsot (pl. Frequencies) a minimum és maximum érték megállapításához. A kapott minimum érték 0,00012, a maximum érték 0,99940.

7. feladat:

Rajzolja ki a ROC görbét! Mit vizsgálunk a ROC görbével? Mekkora a görbe alatti terület nagysága? Végezzünk tesztet arra, hogy a ROC görbe szignifikánsan különbözik-e a 45 fokos egyenestől!

A feladat megoldása:

A ROC görbe a következő menüpont kiválasztásával érhető el:

Analyze → ROC Curve...

A jobboldali 'Test Variable' dobozba helyezzük át a becsült valószínűségeket ('Predicted probability [PRE_1]), a 'State Variable' dobozba pedig a 'Previously defaulted [Default]' változót. A 'Value of State Variable'-t állítsuk 1 értékre, mivel az eredményváltozó 1 értékkel jelölt kategóriájára futtattuk a modellt. A 'Display' dobozban kérjük a referencia vonal jelölését ('With diagonal reference line'), valamint a görbe alatti terület tesztjéhez tartozó standard hibát és konfidencia intervallumot ('Standard error and confidence interval') is.

A ROC is a modell illeszkedését méri. Az x tengely különböző vágási értékek mellett a modell alapján fizetőképtelennek besorolt, de ténylegesen fizetőképes ügyfelek összes, ténylegesen fizetőképes ügyfélhez viszonyított arányát méri. A y tengelyről pedig a különböző vágási értékek mellett a modell alapján fizetőképtelennek besorolt és ténylegesen fizetőképtelen ügyfelek összes, ténylegesen fizetőképtelen ügyfélhez viszonyított arányát olvashatjuk le. A görbe egy-egy pontja azt mutatja meg, hogy bizonyos vágási értékhez milyen aránypárok tartoznak. Minél távolabb helyezkedik el a ROC görbe a 45 fokos (x=y) egyenestől, annál jobban illeszkedik a kapott modell.
Az 1. feladatbeli modell görbe alatti területének (AUC) nagysága 0,856, amely meghaladja a gyakorlatban alkalmazott 0,700 küszöbértéket. Az elvégzett teszt nullhipotézis az, hogy a modell (a konstanson kívül) nem magyaráz semmit. Ebben az esetben a modellel is bizonytalan az ügyfelek besorolása. Ilyenkor a ROC görbe megegyezik a 45 fokos egyenessel, azaz a görbe alatti terület 0,5. Jelen esetben a teszt p értéke alapján minden szokásos szignifikancia szint mellett elvetjük a nullhipotézist, azaz a ROC szignifikánsan különbözik a 45 fokos egyenestől, így a görbe alatti terület a 0,5 értéktől.

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,856	,016	,000	,825	,886

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

8. feladat:

Vizsgálja meg a leverage és a Cook távolság értékeket! Mekkora a leverage és a Cook távolság maximális értéke a fizetőképes és fizetőképtelen ügyfelek körében külön-külön vizsgálva? A leverage értékét tekintve melyek a becslést leginkább befolyásoló pontok?

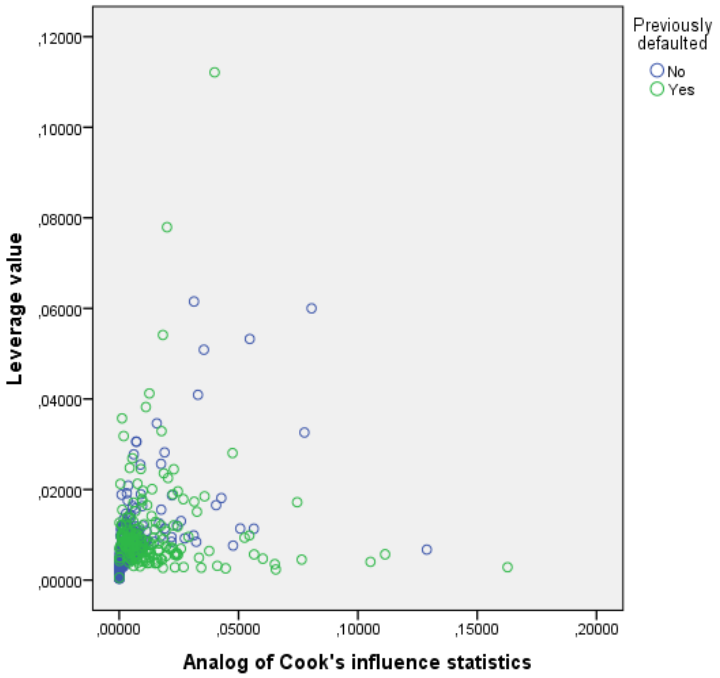
A feladat megoldása:

A feladat megválaszolásához futtassuk újra a modellt a 'Save...' menüpontban a 'Cook's' és a 'Leverage values' lehetőségeket kérve. Ezután az adattáblákban 'LEV_'1 és 'COO_1' változónevek alatt megtalálhatjuk az egyes ügyfelekhez tartozó értéket.
A fizetőképes és fizetőképtelen ügyfelek körében ezek maximális mértékének lekérdését elvégezhetjük az **Analyze → Descriptive Statistics → Explore...**

menüpont alatt. A 'Dependent List' dobozba áthelyezzük a 'LEV_'1 és 'COO_1' változókat, a 'Factor List' dobozba pedig a 'Previously defaulted [default]' változót, valamint a 'Display' dobozban a 'Statistics' lehetőséget kérjük.

A leverage és Cook távolság értékei a regressziós becslést befolyásoló megfigyelések detektálására szolgálnak (lásd még Lineáris regresszió fejezet). Az 1 feletti Cook távolság értékkel rendelkező, illetve 0,2 leverage érték feletti megfigyeléseket kockázatos bevonni a regresszióba, a 0,5 feletti leverage értékkel rendelkezőek bevonása pedig kerülendő, mivel torzítják a regressziós együtthatók becslését. A határértéket átlépő megfigyeléseket célszerű kizárni az elemzésből és újrafuttatni a modellt a torzító pontok nélkül.

A leverage maximális értéke a fizetőképes ügyfelek körében 0,06152, a fizetésképtelen ügyfelek körében 0,11213. A Cook távolság maximális értéke rendre 0,12888, illetve 0,16272. Az egyes ügyfelekhez tartozó Cook távolság értékek sehol sem haladják meg az 1-es küszöbértéket, illetve egyetlen ügyfélhez tartozó leverage érték sem haladja meg a 0,2 küszöbértékét.
Érdemes pontdiagrammal ábrázolni[10] a leverage és Cook távolság értékeket a befolyásoló pontok azonosításához. Az ábrán is láthatjuk, hogy a kockázatos határokat egyetlen ügyfél sem haladja meg, így nincs szükség a regressziós modell újrabecslésére.



9. feladat:
Felléphet-e multikollinearitás logisztikus regresszió esetén? Hogyan vizsgálhatjuk a multikollinearitás jelenlétét az 1. feladatbeli modellben?

A feladat megoldása:

[10] A **Graphs → Legacy Dialogs → Scatter/Dot** menüpontban érhető el Simple Scatter néven.

Logisztikus regresszió esetén is felléphet a magyarázó változók közötti multikollinearitás. Ennek vizsgálatára nem létezik beépített opció a logisztikus regressziónál.
A magyarázó változók közötti multikollinearitásra utalhat a változók közötti páronkénti magas korreláció, illetve a páronkénti lineáris korrelációs együttható magas értéke.

A változók közötti páronkénti korrelációt tartalmazó korrelációs mátrixot kérhetjük az 1. feladatbeli modell ismételt futtatásával az 'Option...' menüpontban a 'Correlations of estimates' lehetőséget kérve.
Az eredmények alapján az 'employ' és a 'creddebt' változók között közepesnél erősebb negatív irányú kapcsolat van.

Correlation Matrix

		Constant	employ	address	debtinc	creddebt
Step 4	Constant	1,000	-,363	-,367	-,676	,316
	employ	-,363	1,000	,073	,074	-,628
	address	-,367	,073	1,000	-,055	-,251
	debtinc	-,676	,074	-,055	1,000	-,400
	creddebt	,316	-,628	-,251	-,400	1,000

A Pearson-féle lineáris korrelációs együttható az **Analyze → Correlate → Bivariate** menüpontban érhető el. Az eredménytáblázat alapján a 'creddebt' és a 'debtinc' változók között van közepes erősségű, pozitív irányú lineáris kapcsolat.

Correlations

		Years with current employer	Years at current address	Debt to income ratio (x100)	Credit card debt in thousands
Years with current employer	Pearson Correlation	1	,345**	-,034	,382**
	Sig. (2-tailed)		,000	,327	,000
	N	850	850	850	850
Years at current address	Pearson Correlation	,345**	1	-,033	,162**
	Sig. (2-tailed)	,000		,337	,000
	N	850	850	850	850
Debt to income ratio (x100)	Pearson Correlation	-,034	-,033	1	,515**
	Sig. (2-tailed)	,327	,337		,000
	N	850	850	850	850
Credit card debt in thousands	Pearson Correlation	,382**	,162**	,515**	1
	Sig. (2-tailed)	,000	,000	,000	
	N	850	850	850	850

**. Correlation is significant at the 0.01 level (2-tailed).

A kapott eredmények alapján érdemes lenne a 'debtinc' változó nélkül újrafuttatni a modellt és összehasonlítani a kapott eredményeket az 1. feladatbeli modellel.

10. feladat:

Végezzen logisztikus regressziós elemzést (Enter módszer) az 'age', 'employ', 'address', 'income', 'debtinc' változók, az 'age' négyzetre emelésével képzett változó, illetve az 'employ' és az 'address' változók interakciójával képzett változó alapján a 'default' változó két kategóriája esetében (cut value: 0,5).
A kapott modell mely változói lesznek szignifikánsak, melyek nem? Értelmezze a négyzetes és interakciós tagok együtthatóit! Röviden értékelje a modell illeszkedését!

A feladat megoldása:

Az 'age' változó négyzetéből képzett változó előállítását a **Transform → Compute Variable...** menüpontban végezhetjük el. A 'Target Variable' dobozban adjuk meg az új változó nevét (pl. 'age_negyz'), a 'Numeric Expression' dobozban pedig az 'age' változó transzformációját (például age * age).
Az 'employ' és 'address' változók keresztszorzatát az **Analyze → Regression → Binary logistic...** menüpontban tudjuk előállítani. A változók baloldali listájában egyszerre kijelöljük a két változót, majd a 'Covariates' dobozba együttesen áthelyezzük őket a nyíl alatti (>a*b> feliratú) gombbal.
A többi magyarázó változó bevonása és a modell beállításai a korábbi feladatokban megismert módon történnek.

A magyarázó változók, illetve keresztszorzat és a négyzetes tag bevonása esetén az 'age', az 'age_negyz' változók és a konstans nem szignifikánsak. Az 'employ' és az 'address' változók ceteris paribus szignifikánsan csökkentik a fizetőképtelenség bekövetkezésének esélyét, együttes hatásuk azonban kismértékben növeli (interakció exp(b) értéke 1,006, p = 0.031) azt. Az 'income' és a 'debtinc' változók ceteris paribus szignifikánsan növelik a fizetőképtelenség bekövetkezésének esélyét.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	age	-,109	,106	1,060	1	,303	,897	,729	1,104
	age_negyz	,002	,001	1,915	1	,166	1,002	,999	1,005
	employ	-,250	,039	41,854	1	,000	,779	,722	,840
	address	-,129	,032	16,215	1	,000	,879	,826	,936
	address by employ	,006	,003	4,673	1	,031	1,006	1,001	1,012
	income	,016	,004	13,924	1	,000	1,016	1,007	1,024
	debtinc	,160	,017	92,542	1	,000	1,174	1,136	1,213
	Constant	,094	1,807	,003	1	,959	1,098		

a. Variable(s) entered on step 1: age, age_negyz, employ, address, address * employ , income, debtinc.

A modell illeszkedése elfogadható. A Hosmer-Lemeshow teszt nullhipotézisét nem tudjuk elvetni (p érték = 0,827), a modell illeszkedésében nem találtunk szignifikáns eltérést az előrejelzett értékektől. A Nagelkerke R négyzet értéke 0,387, nem túl magas, azonban elfogadható érték.
A modell teljes találati aránya 79,7%, a fizetőképesek találati aránya 91,5%, a fizetőképteleneké 46,4%.

Gyakorló feladatok

1. Hasonlítsa össze az 1. és a 10. feladatban illesztett modellek klasszifikációs jóságát a ROC görbe segítségével! Melyik modell jobb a ROC görbe alapján?
2. Bővítse az 1. feladatbeli modellt az 'age' változó négyzetes tagjával, valamint az 'address' és az 'employ' változók interakciójával! A modell illesztése során forward Wald módszert alkalmazzon! Értékelje és értelmezze a kapott modellt (bevont szignifikáns változók és hatásuk, modell illeszkedése, klasszifikáció jósága, torzító pontok detektálása)!

3. Hasonlítsa össze az 1. feladatban, a 10. feladatban, illetve a Gyakorló feladatok 2. feladatában illesztett modellek klasszifikációs jóságát a ROC görbe segítségével! Melyik modell jobb a ROC görbe alapján?

Irodalomjegyzék

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Kovács E., Gray R. [2001]: Az általánosított lineáris modell és biztosítási alkalmazásai.
Statisztikai Szemle 8, p. 689-702.

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. Miért nem alkalmazható a többváltozós lineáris regressziós modell bináris célváltozó esetén?

a) A célváltozó nem normális, hanem bináris eloszlást követ.
b) A célváltozó többváltozós lineáris regressziós modellel becsült értékei a [0,1] intervallumon kívüli értéket is felvehetnek a mintában.
c) Mindkét előző állítás igaz.
d) Egyik előző állítás igaz.

2. Melyik állítás igaz a bináris logisztikus regresszióra?

a) A célváltozó nominális mérésű skálájú.
b) A magyarázó változók között nominális, ordinális vagy magasabb, intervallum és arány skálán mért változók is egyaránt előfordulhatnak.
c) Egy adott csoportba kerülés valószínűségét becsüli.
d) Mindegyik előző válasz helyes.

3. Milyen mutatókkal vizsgálhatjuk a logisztikus regresszió illeszkedésének jóságát?

a) A Hosmer-Lemeshow teszttel, amely esetén nullhipotézis elvetése mutatja a modell jó illeszkedését.
b) Nagelkerke-féle R^2 mutatóval, amely a bevont magyarázó változók által megmagyarázott variancia százalékát méri.
c) A klasszifikáció tábla segítségével, amely a modell által helyesen besorolt megfigyelések arányát méri.
d) Egyik előző válasz sem helyes.

4. Mit mondhatunk a logisztikus modellben egy magyarázó változó hatásáról?

a) Ha a magyarázó változó együtthatója pozitív, az növeli, ha negatív, az csökkenti a vizsgált esemény bekövetkezésének esélyét.
b) A magyarázó változóhoz tartozó exp(b) azt mutatja meg, hogy a magyarázó változó egységnyi változása hányszorosára változtatja a vizsgált esemény bekövetkezésének esélyét.
c) A magyarázó változó parciális hatása nem állandó.
d) Mindegyik előző válasz helyes.

5. Melyik állítás igaz a ROC görbére?

   a) A modellklasszifikáció jóságát méri különböző vágási értékek (cutoff value) mellett.
   b) Az egyes tengelyeken a modell által jól besorolt megfigyelések arányát méri a bináris célváltozó csoportjaiban.
   c) Jó modellilleszkedést mutat, ha a ROC alatti terület szignifikánsan különbözik a nullától.
   d) Mindegyik előző válasz igaz.

6. fejezet

FAKTORELEMZÉS

A módszer rövid összefoglalása

A gyakorlatban nem minden tulajdonság mérhető közvetlenül egy bizonyos mutatószámmal. Valamely országok „infrastrukturális fejlettségi” tulajdonsága például többféle mutatószám együttes értékelésével jellemezhető. Az úgynevezett „látens” (közvetlenül nem mérhető) változók méréséhez nyújthat segítséget a faktorelemzés.

A faktorelemzésben sokféle modell alkalmazható. Egy általános modellfelírás szerint a faktorelemzésben az intervallum vagy arány mérési szintű változók adatait tartalmazó mátrix felírható két másik mátrix összegeként, amelyek közül az egyik a közös faktorok hatását tükrözi, a másik pedig a hibatag mátrix. A faktorelemzés alapegyenlete szerint az elemzésben szereplő változók esetében számolható R korrelációs mátrix felírható a következőképpen: $R = L \cdot L^T + U^2$, ahol L mátrix az úgynevezett faktorsúlyok mátrixa, U^2 pedig a hibatagok diagonális kovarianciamátrixa.

A faktorelemzés matematikai hátterében nagy szerepe van bizonyos mátrixok sajátérték-sajátvektor felbontásának. A faktorelemzés általános modelljében az U^2 mátrix ismeretében $L \cdot L^T$ mátrix (a redukált korrelációs mátrix) sajátérték-sajátvektor felbontásával számolhatók eredmények. Ha az $L \cdot L^T$ mátrix helyett az R korrelációs mátrix sajátérték-sajátvektor felbontásával kerül sor az eredmények számolására egy faktorelemzéshez ettől eltekintve hasonló elemzésben, akkor főkomponens-elemzésről van szó. A főkomponens-elemzés modellje egyfajta faktormodellnek tekinthető (*Hajdu* [2003], 386. oldal). A főkomponens-elemzés eredményei között faktorok helyett komponensek említése szerepel, és a faktorelemzés $R = L \cdot L^T + U^2$ alapegyenlete helyett a főkomponens-elemzésben $R = C \cdot C^T$ összefüggést szokás felírni, ahol C mátrix az úgynevezett komponens-súlyok mátrixa.

Az összesen létrehozható faktorok (illetve komponensek) közül mindössze az elemzés céljának megfelelően magas sajátértékekhez kapcsolódóakat szokás „kivonni” (vagyis lényegében több más eredmény, például a reprodukált korrelációs mátrix számolásához alkalmazni). A faktorelemzésben többek között például a Bartlett-féle teszt illetve a KMO-értékek alapján lehet megállapítani, hogy az adatok megfelelőek-e a faktorelemzés végzésére. A számolt eredmények megfelelősége a „kivont” faktorok (illetve komponensek) által együttesen magyarázott variancia, a kommunalitás értékek, valamint a faktormátrix (illetve a komponens mátrix) tartalma alapján értékelhető. A faktorok (illetve komponensek) értelmezését rotálás (a faktorok illetve komponensek forgatása) is elősegítheti.

Megoldási módszerek és az eredmények értelmezése

Mivel a faktorelemzés, illetve a főkomponens-elemzés során az eredmények valamely mátrix sajátérték-sajátvektor felbontásából adódnak, így a következőkben a számolások matematikai hátterével, illetve egyes matematikai összefüggések értelmezésével is foglalkozunk.

A gyakorló feladatoknál említett változók a telco.sav adatai között találhatók.

1. feladat:

Hasonlítsa össze a „megőrzött” („extracted”) komponensek számát a „longmon”, „longten”, „cardmon” és „cardten” változók alapján végzett főkomponens-elemzésben a „pager” változó két csoportjában azoknál a megfigyeléseknél, ahol a „wireless” változó értéke nem nulla! (A komponensek közül az egynél nagyobb sajátértékűeket vonja ki az elemzésben.)

A feladat megoldása:

A feladat megoldása során először a szűrési feladattal foglalkozunk, a következő menüpont választásával:

Data → Select Cases …

A menüpont kiválasztása után megjelenő ablakban a „Select” feliratnál az „If condition is satisfied” feliratnál az „If…” gombra kattintás után megjelenő újabb ablakban található egy – a szűrési feltételek definiálásához kapcsolódó – képletek beírására alkalmas rész, ahová a bináris „wireless” változóra vonatkozóan a következő feltétel írható be:

wireless=1

A feladatleírás szerint a főkomponens-elemzést két külön csoportban is el kell végezni, ami megvalósítható lenne úgy is, hogy a szűrési feltételeket külön-külön módosítjuk a két feladat-részben, illetve megoldás lehet a következő menüpont választása után megjelenő ablakban egy megfelelő beállítás választása is:

Analyze → Dimension Reduction → Factor …

A menüpont választása után megjelenő ablakban a „Selection Variable:” felirathoz a „pager” változó kerül, a „Value …” gombra kattintás után megjelenő ablakban pedig a „Value for Selection Variable:” felirat után a bináris „pager” változó két kategóriájára vonatkozóan először 0, majd 1 érték kerül a két különböző főkomponens-elemzésben. Ezután a „Continue” gomb megnyomásával vissza lehet térni a főkomponens-elemzés többi beállítási lehetőségéhez.

Total Variance Explained[a]

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,842	71,048	71,048	2,842	71,048	71,048
2	,988	24,695	95,742			
3	,152	3,790	99,533			
4	,019	,467	100,000			

Extraction Method: Principal Component Analysis.

a. Only cases for which Paging service = No are used in the analysis phase.

Az elemzésben szereplő változókat a „Variables:” feliratnál lehet elhelyezni. Az „OK” gomb megnyomásával számolható eredmények között megtalálható az előző táblázat

is, amelyből megállapítható, hogy az elemzésben egyetlen „kivont” komponens van abban a csoportban, ahol a „pager” változó értéke nulla.

A másik csoportban (ahol a “pager” változó értéke 1) az elemzésben két komponens “kivonására” kerül sor, ahogyan azt a következő táblázat is mutatja:

Total Variance Explained[a]

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,680	67,007	67,007	2,680	67,007	67,007
2	1,175	29,368	96,375	1,175	29,368	96,375
3	,129	3,213	99,588			
4	,016	,412	100,000			

Extraction Method: Principal Component Analysis.

a. Only cases for which Paging service = Yes are used in the analysis phase.

A főkomponens-elemzésben a korrelációs mátrix sajátértékei a komponensek varianciáival egyeznek meg. A komponensek közül mindössze az egynél nagyobb sajátértékkel rendelkezők kivonására került sor. Ezt például azzal lehet indokolni, hogy az elemzésben szereplő eredeti változók esetében a sztenderdizált változók varianciája (ami egyben a korrelációs mátrix főátlójának egy eleme is) egységnyi, és ehhez az értékhez lehet hasonlítani a komponensek varianciáit is.

Ha valamely adathalmaz egyes csoportjaiban jelentősen eltérnek a főkomponens-elemzéssel kapott eredmények, akkor az egyes csoportok főkomponens-elemzéses eredményeit érdemes külön tanulmányozni. A következőkben ennek megfelelően ebben a fejezetben a „pager” változó két csoportjának adatai külön elemzésekben szerepelnek.

2. feladat:

Az 1. feladat adatai alapján mekkora a korrelációs mátrix determinánsa abban az adathalmazban, ahol a „pager” változó értéke nulla? Hogyan értelmezhető ez az eredmény?

A feladat megoldása:

A korrelációs mátrix determinánsának számolásához az 1. feladatban alkalmazott beállításokat kiegészítjük: a „Descriptives …” gomb megnyomása után megjelenő ablakban a „Correlation Matrix” felirat alatt található „Determinant” lehetőséget is választjuk, valamint érdemes még ezenkívül a „Coefficients” lehetőséget is választani, hogy a korrelációs mátrix értékei is megtalálhatók legyenek a számolt eredmények között.

A korrelációs mátrix determinánsának értelmezéséhez röviden érdemes foglalkozni a mátrix determináns számolás témájával is. A mátrix determináns számolás a matematika több területén is kiemelkedő jelentőségű. Ezt mutatja például, hogy ha egy négyzetes mátrix (amelynek ugyanannyi sora van mint oszlopa) determinánsa nem nulla, akkor létezik az inverz mátrix. A legegyszerűbb esetben, amikor egy négyzetes

mátrixnak két oszlopa van és az elemeket a mátrixban *a, b, c* és *d* jelöli, akkor a mátrix determinánsa:

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = a \cdot d - b \cdot c$$

Tekintsük például a korrelációs mátrix esetét akkor, ha az elemzésben két változó szerepel. A korrelációs mátrix főátlójában mindegyik elem 1. Jelölje a korrelációs mátrixot *R*, a két változó közötti (lineáris, Pearson-féle) korrelációs együtthatót pedig *r*, ekkor a korrelációs mátrix determinánsa:

$$\det(R) = \det\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} = 1 - r^2$$

Könnyen belátható, hogy ha a két változó közötti korrelációs együttható nulla, akkor a korrelációs mátrix determinánsa 1, ha pedig a korrelációs együttható +1 vagy -1, akkor a korrelációs mátrix determinánsa nulla. A főkomponens-elemzéshez az adatok általában akkor megfelelők, ha az elemzésben szereplő változók között erősek a lineáris összefüggések. (Pontosabban fogalmazva ezenkívül például még a parciális korrelációs együtthatók értékei is fontosak az eredmények „jóságának" megítélésénél, de a Pearson-féle lineáris korrelációs együtthatók értékeiből is sok következtetés adódik.) Két változónál tehát akkor lehet a főkomponens-elemzésnél jó eredményekre számítani, ha a két változó közötti lineáris korrelációs együttható értéke abszolút értékben magas. A korrelációs mátrix determinánsa nulla és egy közötti érték, a maximum egységnyi érték akkor mérhető, ha a korrelációs mátrix egységmátrix. A főkomponens-elemzéshez az adatok annál inkább megfelelőek, minél kisebb a korrelációs mátrix determinánsa.

Ebben a feladatban az eredmények között megtalálható a korrelációs mátrix és a determinánsa (amelyet a következő táblázat alatti bekarikázott érték jelöl):

Correlation Matrix[a,b]

		Long distance last month	Long distance over tenure	Calling card last month	Calling card over tenure
Correlation	Long distance last month	1,000	,979	,270	,651
	Long distance over tenure	,979	1,000	,282	,688
	Calling card last month	,270	,282	1,000	,728
	Calling card over tenure	,651	,688	,728	1,000

a. Only cases for which Paging service = No are used in the analysis phase.

b. Determinant = ,008

Az eredmények között szereplő determináns érték 0,008, ami meglehetősen közel van a korrelációs mátrix determináns-értékének elméleti minimumához (a nulla értékhez), vagyis e mutatószám alapján az elemzésben található adatok a főkomponens-elemzéshez megfelelőnek tekinthetők.

3. feladat:

Az 1. feladat adatai alapján mekkora a KMO érték és a Bartlett-féle khi-négyzet teszthez kapcsolódó empirikus szignifikanciaszint abban az adathalmazban, ahol a „pager" változó értéke egységnyi? Hogyan értelmezhetők ezek az eredmények?

A feladat megoldása:

A KMO érték és a Bartlett-féle khi-négyzet teszthez kapcsolódó empirikus szignifikanciaszint (p-érték) számolásához az 1. feladatban alkalmazott beállításokat kiegészítjük: a „Descriptives" gomb megnyomása után megjelenő ablakban a „Correlation Matrix" felirat alatt a „KMO and Bartlett's test of sphericity" lehetőséget is választjuk. Ezután a „Continue" majd az „OK" gomb megnyomása után az eredmények között megtalálható a következő táblázat is:

KMO and Bartlett's Test[a]

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,516
Bartlett's Test of Sphericity	Approx. Chi-Square	1031,515
	df	6
	Sig.	,000

a. Only cases for which Paging service = Yes are used in the analysis phase.

A táblázatban szereplő (bekarikázással jelölt) eredmények alapján a feladatban a KMO érték 0,516, a Bartlett-féle khi-négyzet teszthez kapcsolódó p-érték pedig nullához közeli érték (három tizedesjegyre kerekítve 0,000). Ez a két érték arra utal, hogy az elemzésben szereplő adatok a főkomponens-elemzéshez megfelelőnek tekinthetők.

A KMO érték és a Bartlett-féle khi-négyzet teszt eredményeinek értelmezéséhez érdemes röviden foglalkozni a számolások részleteivel is. A főkomponens-elemzésnél a Bartlett-féle khi-négyzet teszt (*Bartlett's test of sphericity*) nullhipotézise az, hogy a korrelációs mátrix egységmátrix (ezenkívül ez a teszt arra a feltevésre is épül, hogy a változók eloszlása többdimenziós normális eloszlás), így ha ennél a tesztnél a p-érték (amit a táblázatban a „Sig." felirat jelöl) nulla közeli érték, akkor a korrelációs mátrix nem tekinthető egységmátrixnak, ami a főkomponens-elemzés eredményei szempontjából kedvező.

A KMO érték számolása során a (Pearson-féle) lineáris korrelációs együtthatókon kívül a parciális korrelációs együtthatóknak is szerepe van (a KMO érték számolásának képlete megtalálható például *Kovács* [2011] 95. oldalon). Ha a parciális korrelációs együtthatók mindegyike nulla lenne, akkor a KMO érték egységnyi lenne, az abszolút értékben nullánál nagyobb parciális korrelációs együtthatók pedig csökkentik a KMO mutatószám értékét. Egy lehetséges értékelés szerint (*Kovács* [2011], 95. oldal) 0,9 feletti KMO érték arra utal, hogy az adatok kiválóan megfelelnek a főkomponens-elemzéshez, míg 0,5 alatti KMO érték alapján a főkomponens-elemzés nem alkalmazható adatelemzési módszerként.

A feladatban szereplő eredmények alapján az adatok főkomponens-elemzéshez alkalmazhatónak tekinthetők, mivel a KMO-érték 0,5 feletti és a Bartlett-féle khi-

négyzet teszt p-értéke is kisebb, mint a gyakran alkalmazott 5 százalékos érték (az eredmények azzal együtt érdekesek, hogy a Bartlett-féle khi-négyzet teszthez kapcsolódó, a többdimenziós normális eloszlásra vonatkozó előfeltevés tesztelésével a feladat megoldása során nem foglalkoztunk).

4. feladat:

A 2. feladat adatai alapján számolt komponens mátrix tartalma alapján lehetséges kiszámolni a „cardmon" változó kommunalitás-értékét?

A feladat megoldása:

A főkomponens-elemzésben az *R* korrelációs mátrix a sajátérték-sajátvektor felbontás figyelembevételével felírható $R = A \cdot \Lambda \cdot A^T$ egyenlettel, ahol *Λ* diagonális mátrix főátlójában a korrelációs mátrix sajátértékei találhatók, az *A* mátrix pedig az „egységnyi hosszúságú" sajátvektorokra utal. A nem „egységnyi hosszú" sajátvektorok is számolhatók, például $c_1 = a_1 \cdot \sqrt{\lambda_1}$ módon. A c_1 vektor elemei az első komponens és a változók közötti korrelációs együtthatóként értelmezhetők. A $c_1, ..., c_p$ vektorokat a *C* mátrixban foglalhatjuk össze (amelynek neve komponens mátrix). A komponens mátrix alapján meghatározható a változók esetében a kommunalitás értéke is. A 2. feladat adatai alapján számolható eredmények között található a következő táblázat, amelyből megállapítható, hogy a „cardmon" változó esetében (három tizedesjegyre kerekítve) 0,395 a kommunalitás-érték:

Communalities[a]

	Initial	Extraction
Long distance last month	1,000	,801
Long distance over tenure	1,000	,828
Calling card last month	1,000	,395
Calling card over tenure	1,000	,818

Extraction Method: Principal Component Analysis.

a. Only cases for which Paging service = No are used in the analysis phase.

A 2. feladat adatai alapján számolható eredmények között szintén megtalálható a komponens mátrix, amelynek alapján az első komponens és a "cardmon" változó közötti lineáris korrelációs együttható értéke 0,628. Mivel ebben a feladatban mindössze egyetlen komponens "kivonására" került sor, így a "cardmon" változó kommunalitás-értéke és e korrelációs együttható között viszonylag egyszerű az összefüggés: a kommunalitás-érték a komponens-mátrixban található érték négyzete: $0{,}395 = 0{,}628^2$ (figyelembe véve hogy az eredményeket tartalmazó táblázatokban három tizedesjegyre kerekített értékek vannak).

Component Matrix[a,b]

	Component
	1
Long distance last month	,895
Long distance over tenure	,910
Calling card last month	,628
Calling card over tenure	,904

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

b. Only cases for which Paging service = No are used in the analysis phase.

Érdemes azt is megemlíteni, hogy egynél több kivont komponens esetében kissé bonyolultabbá válik az a képlet, amelynek alapján a komponens mátrix tartalmából a kommunalitás-értékek számolhatók.

5. feladat:

Mekkora a 2. feladatban a sajátértékek összege?

A feladat megoldása:

A főkomponens-elemzésnél az egyik cél, hogy az első komponens varianciája maximális legyen. E feladat megoldása során az az eredmény adódik, hogy az első komponens varianciája az első (legnagyobb) sajátértékkel egyezik meg. Ahogyan az az előző feladat megoldásában is szerepel, a sajátértékeket tartalmazó diagonális mátrixot jelölheti Λ, ahol a korrelációs mátrix sajátérték-sajátvektor felbontásával összefüggésben $A^T \cdot R \cdot A = \Lambda$. A Λ mátrix főátlójában lévő elemek összege a főkomponens-elemzés eredményeinek matematikai levezetésével összefüggésben megegyezik a korrelációs mátrix főátlójában lévő elemek összegével. Az *R* korrelációs mátrix esetében a főátlóban lévő elemek összege a változók számával egyezik meg, ha tehát a főkomponens-elemzésben az elemzés alapja a korrelációs mátrix, akkor a sajátértékek összege megegyezik a változók számával. Ebben a feladatban a sajátértékek összege 4.

6. feladat:

Az összes variancia mekkora részét hordozzák a kivont komponensek a 3. feladatban?

A feladat megoldása:

A 3. feladatban is négy változó szerepel, az 5. feladat megoldását is figyelembe véve tehát a komponensek varianciáinak összege is 4. Az első két (kivont) komponenst (amelyeknek varianciája egynél nagyobb) figyelembe véve a kivont komponensek

összesen a variancia $\frac{2{,}680+1{,}175}{4}=0{,}96375$ részét hordozzák. Ezt az eredményt a következő táblázat (bekarikázással jelölt) értékei is mutatják:

Total Variance Explained[a]

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,680	67,007	67,007	2,680	67,007	67,007
2	1,175	29,368	96,375	1,175	29,368	96,375
3	,129	3,213	99,588			
4	,016	,412	100,000			

Extraction Method: Principal Component Analysis.

a. Only cases for which Paging service = Yes are used in the analysis phase.

7. feladat:

Alkalmazzon „varimax” rotálást a 3. feladat adatai esetében! A „varimax” rotálás hatására megváltozik a kivont komponensek által magyarázott összes variancia értéke?

A feladat megoldása:

A rotálási beállításokat a 3. feladat beállításainak kiegészítéseképpen a „Rotation …” gomb megnyomásával lehet megtekinteni, és a „Method” feliratnál a „Varimax” lehetőség bejelölésével kérhető a „varimax” rotálási módszer alkalmazása. Az eredmények között található a következő táblázat is, amelyben a (bekarikázással jelölt) értékek mutatják, hogy a kivont komponensek által magyarázott összes variancia értéke nem változik a „varimax” rotálás hatására:

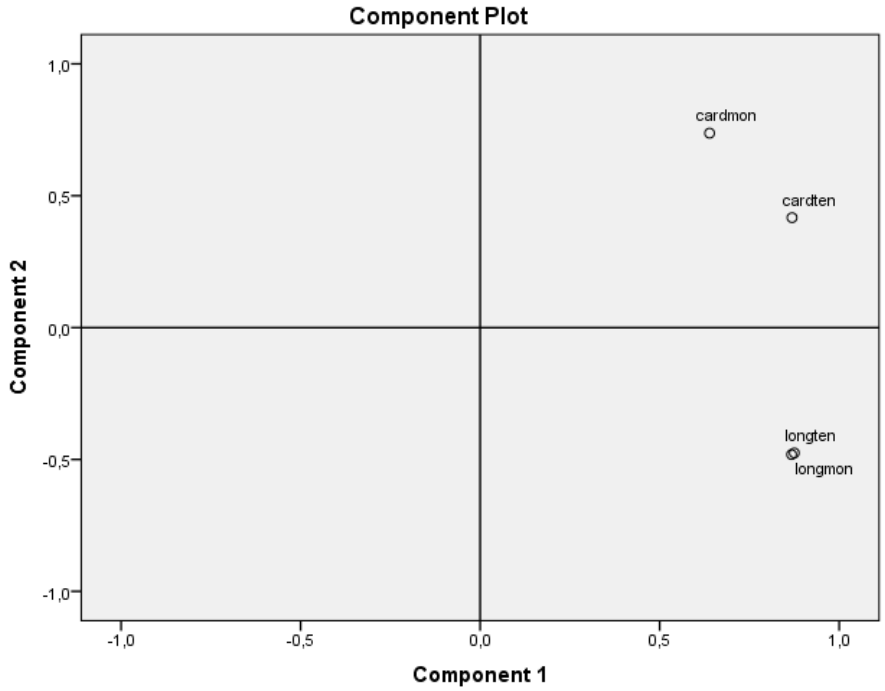
Total Variance Explained[a]

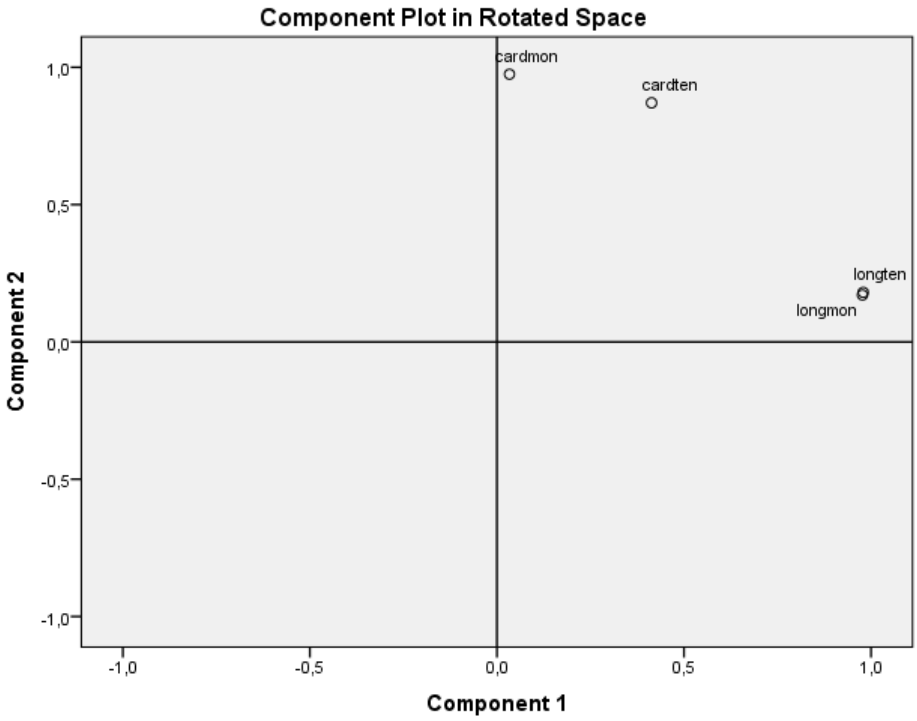
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,680	67,007	67,007	2,680	67,007	67,007	2,085	52,131	52,131
2	1,175	29,368	96,375	1,175	29,368	96,375	1,770	44,244	96,375
3	,129	3,213	99,588						
4	,016	,412	100,000						

Extraction Method: Principal Component Analysis.

a. Only cases for which Paging service = Yes are used in the analysis phase.

A rotálás (forgatás) elősegítheti a faktorok, illetve komponensek értelmezését, mivel ennek hatására a komponens mátrix tartalma is változhat, és így könnyebb lehet megállapítani, hogy mely (rotált) faktorok (illetve komponensek) és eredeti változók között vannak jelentős összefüggések. Ha az ebben a feladatban alkalmazott főkomponens-elemzési beállításokat tovább bővítve a „Rotation …” gomb megnyomása után megjelenő ablakban a „Display” feliratnál a „Loading plot(s)” lehetőséget is bejelöljük, a két oszlopot tartalmazó komponens mátrix tartalmát grafikusan is meg lehet jeleníteni, ahogyan ezt a következő két ábra is mutatja (a két ábra a rotálás előtti és rotálás utáni eredményeket szemlélteti):





A két ábra összehasonlításával megállapítható, hogy a rotálás utáni helyzetben az eredeti változók koordinátái többnyire „közelebb” kerültek az egyes tengelyekhez, ami a tengelyek értelmezését is egyszerűsíti.

8. feladat:
Hogyan értelmezhető a 7. feladatban a rotálás után az első komponens?

A feladat megoldása:

A komponensek értelmezése elsősorban a komponens mátrix alapján történhet. A 7. feladatban az eredmények között megtalálható a rotált komponens mátrix is (az egyes változókhoz tartozó sorokban a maximális értéket bekarikázás jelöli):

Rotated Component Matrix[a,b]

	Component	
	1	2
Long distance last month	,977	,171
Long distance over tenure	,979	,181
Calling card last month	,033	,975
Calling card over tenure	,413	,870

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

b. Only cases for which Paging service = Yes are used in the analysis phase.

Mivel a komponens mátrixban található értékek korrelációs együtthatóként értelmezhetők, így a komponensek értelmezésekor azt kell megállapítani, hogy a komponens mátrixban mely változók esetében magasak abszolút értékben az adott oszlopokban található értékek. Ebben a feladatban a rotált első komponenshez tartozó oszlopban a "longmon" és a "longten" változóknak magas a korrelációs együtthatója, tehát (a változók teljes nevét is figyelembe véve) a "long distance" tulajdonsággal valamilyen kapcsolatban lévő mutatószámnak tekinthető az első komponens.

9. feladat:
A 2. feladat adatai alapján végezzen faktorelemzést Principal Axis Factoring módszerrel. Mennyi az egynél nagyobb sajátértékkel rendelkező faktorok száma?

A feladat megoldása:

A főkomponens-elemzés helyett másik faktorelemzési modell kiválasztására az „Extraction …" gomb megnyomása után megjelenő ablakban van lehetőség: a „Method" feliratnál a „Principal components" helyett ebben a feladatban a „Principal axis factoring" lehetőséget választjuk. Az eredmények között megtalálható a komponens mátrixhoz hasonló faktormátrix, ami a faktorsúlyokat tartalmazza:

Factor Matrix[a,b]

	Factor
	1
Long distance last month	,888
Long distance over tenure	,922
Calling card last month	,474
Calling card over tenure	,836

Extraction Method: Principal Axis Factoring.

a. 1 factors extracted. 7 iterations required.

b. Only cases for which Paging service = No are used in the analysis phase.

Ebben a feladatban tehát egyetlen faktorhoz tartozó sajátérték volt nagyobb egynél, így mindössze egy faktor "kivonására" került sor.

Gyakorló feladatok

1. A „longmon", „longten", „tollmon", „tollten", „cardmon" és „cardten" változók alapján végzett főkomponens-elemzésben mekkora a legkisebb kommunalitás-érték?

2. A *Gyakorló feladatok* 1. feladatának adatai alapján végzett főkomponens-elemzésben érdemes lenne valamelyik változót kihagyni az elemzésből?

3. Mekkora a *Gyakorló feladatok* 1. feladatában az első és második komponens közötti (lineáris) korrelációs együttható értéke?

4. A *Gyakorló feladatok* 1. feladatában hogyan értékelhető az adatok megfelelősége a KMO-érték alapján?

5. A *Gyakorló feladatok* 1. feladatának adatai alapján végzett főkomponens-elemzésben mentse el az egynél nagyobb varianciájú komponensek értékeit! Mekkora ennek az elmentett új változónak az átlagos értéke?

6. Végezzen a *Gyakorló feladatok* 1. feladatában szereplő változók alapján faktorelemzést Principal Axis Factoring (PAF) módszerrel! Hasonlítsa össze a reprodukált korrelációs mátrix tartalmát a *Gyakorló feladatok* 1. feladatánál számolható eredményekkel!

7. Hogyan változik a főkomponens-elemzésben a korrelációs mátrix determinánsa, ha a sztenderdizálás nélküli eredeti adatok helyett sztenderdizált változók alapján kerül sor az elemzésre?

8. Hogyan értelmezhetők az „anti-image” korrelációs mátrix főátlójában szereplő értékek?

9. Mekkora valamely változó kommunalitása, ha egy főkomponens-elemzésben az összes komponens „kivonására” sor kerül?

10. Tegyük fel, hogy egy főkomponens-elemzésben az összes komponens értékeinek elmentésére sor került. Az ezen új elmentett változók alapján számolt (a Pearson-féle korrelációs együtthatókat tartalmazó) korrelációs mátrixban elméletileg lehetnek negatív értékek?

11. Elméletileg a Principal Axis Factoring (PAF) módszer alkalmazása esetében a faktorokhoz tartozó sajátértékek között lehet negatív érték?

Irodalomjegyzék

Hajdu Ottó [2003]: *Többváltozós statisztikai számítások*
Központi Statisztikai Hivatal

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. A faktorelemzésben a Principal Axis Factoring módszer alkalmazásakor a reprodukált korrelációs mátrix esetében

   a) a főátlóban minden esetben egységnyi értékek találhatók
   b) a főátlóban szerepelhet bármilyen valós szám, amely abszolútértékben kisebb vagy egyenlő mint egy
   c) a főátlón kívüli elemek pozitív értékek
   d) egyik előző állítás sem helyes.

2. A főkomponens-elemzésben a maximálisan képezhető komponensek száma:

   a) annyi mint a változók száma
   b) kevesebb mint a változók száma
   c) több mint a változók száma
   d) végtelen.

3. Ha két változó közötti korreláció értéke nulla, akkor a két változó alapján végzett főkomponens-elemzésben az első komponens varianciája

   a) 1
   b) -1
   c) nulla
   d) bármennyi lehet.

4. Ha egy korrelációs mátrix minden eleme egyhez közeli érték, akkor a korrelációs mátrix determinánsa megközelítőleg

   a) nulla
   b) 1
   c) -1
   d) egyik előző állítás sem helyes.

5. Egy 4 változót tartalmazó főkomponens-elemzésben a sajátértékek összege

   a) 4
   b) 16
   c) 2
   d) egyik előző válasz sem helyes.

7. fejezet

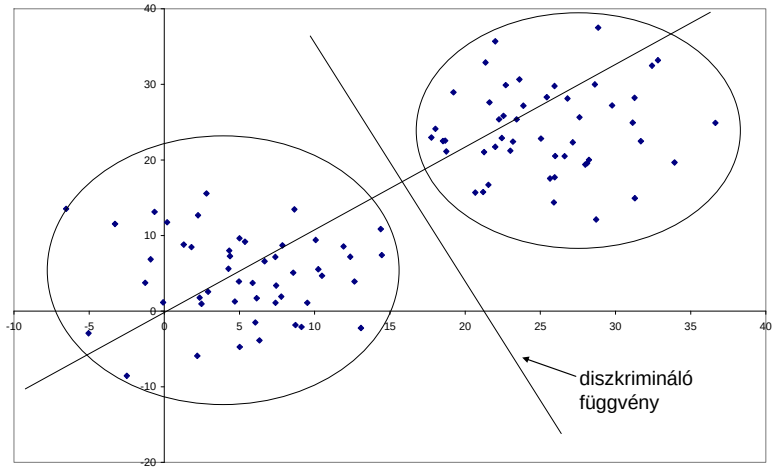
DISZKRIMINANCIA-ELEMZÉS

A módszer rövid összefoglalása

Egyes elemzéseknél szükség lehet a megfigyelések (elemek) csoportokba sorolására (például hitelkérelmek elbírálásánál, hogy kaphat-e hitelt a hiteligénylő vagy nem). A csoporttagság „előrejelzésére" alkalmas többváltozós statisztikai modellek egyik fajtája a diszkriminancia-elemzés. A diszkriminancia-elemzés alapfeladata olyan diszkrimináló függvények meghatározása az eredeti változók lineáris kombinációiként ($y = X \cdot c$, $c^T \cdot c = 1$ feltétel figyelembevételével, ahol az eredeti változók értékeit az n sorral és p oszloppal rendelkező X mátrix tartalmazza), hogy az elemzésben szereplő csoportokat a lehető legjobban el lehessen különíteni a kanonikus térben. A továbbiakban ebben a fejezetben a megfigyelések számát n (illetve a csoportokban n_i), az eredeti változók számát p, az elemzésben szereplő csoportok számát pedig g jelöli.

Az elemzéshez tartozó két fontos alkalmazási előfeltevés, hogy a csoportokban az elemzésben szereplő változók kovarianciamátrixa azonosnak tekinthető, illetve hogy a változók együttes eloszlása többdimenziós normális eloszlás. A feladat megoldása során fontos szerepe van $T = X^T \cdot X = K + B$ mátrixnak, ahol $B = \sum_{i=1}^{g} (n_i - 1) \cdot S_i$, mivel a diszkriminancia-elemzésnél az elemzés céljának matematikai megfogalmazása: $\frac{c^T \cdot K \cdot c}{c^T \cdot B \cdot c} \to \max_c$, ahol a feladat megoldása: $\left(B^{-1} \cdot K - \lambda \cdot E\right) \cdot c = 0$. Ahogyan az ebből a megoldásból is látszik, a diszkriminancia-elemzés során sajátértékek számolására is sor kerül, a sajátértékek, illetve az ezekhez tartozóan számolható diszkrimináló függvények maximális száma pedig $B^{-1} \cdot K$ mátrix rangjával egyezik meg: $\min\{g-1, p\}$.

Például két változó és két csoport esetében a diszkrimináló függvények maximális száma 1, ugyanis: $\min\{g-1, p\} = \min\{2-1, 2\} = 1$. Ezt a helyzetet illusztrálja a következő ábra:



A diszkriminancia-elemzés eredményeinek megfelelőségét többféle mutatószámmal is lehet mérni. Ezek közül például a diszkrimináló függvényekhez tartozó Wilks-lambda és a kanonikus korrelációs értékek összefüggnek $B^{-1} \cdot K$ mátrix sajátértékeivel. Jelölje λ_j ($j=1,...,k$) a $B^{-1} \cdot K$ mátrix sajátértékeit, ahol $k = \min\{g-1, p\}$. Az összes diszkrimináló függvényhez tartozóan számolható Wilks-lambda érték ekkor $\prod_{j=1}^{k} \frac{1}{1+\lambda_j}$, a j-edik diszkrimináló függvényhez tartozóan számolható kanonikus korreláció értéke pedig $\sqrt{\frac{\lambda_j}{1+\lambda_j}}$.

A kanonikus térben a csoportok elkülönülése jobb, ha a Wilks-lambda érték(ek) alacsonyabb(ak), illetve a kanonikus korrelációs érték(ek) magasabb(ak).

Megoldási módszerek és az eredmények értelmezése

A diszkriminancia elemzésben egy csoportosítást definiáló változó és *p* darab (intervallum vagy arány mérési szintű) „magyarázó" változó szerepel. Számos más adatelemzési módszerhez hasonlóan a diszkriminancia-elemzésben is lehetőség van „stepwise" módszer alkalmazására, amikor az elemzésben mindössze a valamilyen szempont alapján szignifikáns hatásúnak tekinthető magyarázó változók kerülnek be a modellbe. A következőkben a feladatok megoldása során a Wilks lambda elven alapuló stepwise módszert alkalmazzuk. A különböző lépésenkénti változószelekciós (stepwise) módszerek jellemzőiről bővebben például *Kovács* [2011] 132-134. oldalán lehet olvasni.

A gyakorló feladatok megoldásánál említett változók a telco.sav adatai között találhatók.

1. feladat:

Végezzen diszkriminancia-elemzést a „longmon", „tollmon", „equipmon", „cardmon" és „wiremon" változók (magyarázó változók), valamint a „custcat" változó (függő változó) alapján stepwise (Wilks' lambda, „Use probability of F") módszerrel azon adatok esetében, amelyeknél egyidejűleg teljesülnek a következő feltételek: a „custcat" változó értéke 4-nél kisebb, az „equip" és a „tollfree" változó értéke pedig egyaránt 1. Teljesülnek az elemzés alkalmazási előfeltevései?

A feladat megoldása:

Érdemes megemlíteni, hogy az adathalmazban az adatok szűrése előtt a „custcat" változónak 4 kategóriája van, amelyeket az 1, 2, 3 és 4 értékek jelölnek, az „equip" és a „tollfree" változók pedig bináris változók. A feltételeknek megfelelő megfigyeléseket a következő menüpont választásával, szűréssel lehet kiválasztani:

Data → Select Cases …

A menüpont kiválasztása után megjelenő ablakban a „Select" feliratnál az „If condition is satisfied" feliratnál található „If…" gombra kattintva megjelenik egy újabb ablak, amelyben található egy – a szűrési feltételek definiálásához kapcsolódó – képletek beírására alkalmas rész, amelybe a következő képlet írható be:

custcat < 4 & equip=1 & tollfree=1

A diszkriminancia-elemzés a következő menüpont választásával kezdhető el:

Analyze → Classify → Discriminant …

A menüpontra kattintás után megjelenő ablakban lehet beállítani, hogy mely csoportok és melyik „magyarázó" változók szerepelnek az elemzésben. A „custcat" változót a „Grouping Variable:" felirathoz, a feladatban szereplő magyarázó változókat pedig az „Independents:" felirathoz helyezzük el. A csoportok meghatározásánál szükség van még a csoportok pontosabb beazonosítására is: a „Define Range …" gombra kattintással ebben az esetben a „Minimum:" felirathoz 1-es, a „Maximum:" felirathoz pedig 3-as érték kerül (figyelembe véve hogy a szűrés után az elemzésben a „custcat" változónak e három kategóriája található meg az adathalmazban).

A stepwise módszer beállításához az „Independents:" felirat alatt található „Use stepwise method" lehetőség bejelölése esetén a „Method…" gomb megnyomása után megjelenő ablakban lehet kiválasztani a lépésenkénti változószelekcióhoz kapcsolódó módszert. Ebben a feladatban a „Criteria" feliratnál a „Use probability of F" lehetőséget választjuk, majd a „Continue" gomb megnyomásával a további beállításokkal foglalkozunk.

A diszkriminancia-elemzés két fontos alkalmazási előfeltevése közül a kovariancia-mátrixok egyezőségének teszteléséhez a Box-M mutatószámon alapuló F-eloszlású tesztstatisztika kapcsolódik, amelynek kiszámításához a „Statistics …" gomb megnyomása után megjelenő ablakban a „Descriptives" feliratnál a „Box's M" lehetőséget jelöljük be. A „Continue" gomb, majd az „OK" gomb megnyomásával számolható eredmények közül a Box-M mutatószámot a következő táblázat tartalmazza:

Test Results

Box's M		10,568
F	Approx.	1,608
	df1	6
	df2	6851,509
	Sig.	,140

Tests null hypothesis of equal population covariance matrices.

A csoportok kovariancia mátrixainak egyezőségére vonatkozó nullhipotézishez tartozó empirikus szignifikancia-szint (p-érték, amelyet a "Sig." feliratú sorban a táblázatban bekarikázás jelöl) 0,140, ami nagyobb mint a statisztikai tesztekben szignifikanciaszintként gyakran alkalmazott 5 százalék (azaz 0,05), vagyis a csoportok kovariancia mátrixai ebben a feladatban egyezőnek tekinthetők.

A diszkriminancia-elemzés másik fontos alkalmazási előfeltevése, hogy a csoportokban az elemzésben szereplő változók együttes eloszlása többdimenziós normális eloszlás. A következőkben az SPSS program alkalmazásával mindössze a változók egydimenziós normális eloszlását lehet tesztelni. Az ezzel kapcsolatos eredmények olyan szempontból tekinthetők relevánsnak a feladatmegoldásban, hogy ha az egyes változók eloszlása nem lenne egydimenziós normális eloszlású, akkor kizárható lenne a változók együttes eloszlására vonatkozóan a többdimenziós normális eloszlás esete. A többdimenziós normális eloszlás tesztelése összetett feladat, és többféle tesztstatisztika is számolható ezzel kapcsolatban, például ehhez kapcsolódik a *McNeil et al.* [2005] (69-70. oldal) által is említett Mardia többváltozós normalitás tesztje (a teszt leírását például *Mardia* [1970] is tartalmazza).

Felmerül a kérdés, hogy mely változók kerültek be a stepwise módszer alkalmazása során az elemzésbe. Erre a kérdésre az eredmények között szereplő struktúra mátrixban is találhatók adatok. A struktúra mátrix a "magyarázó" változók és a (sztenderdizált) kanonikus diszkrimináló függvények értékei közötti egyfajta "kevert" ("pooled") korrelációs együttható értékeket mutatja. A feladatban számolt eredmények alapján megállapítható, hogy a stepwise változószelekciós módszer alkalmazásával a "tollmon" és a "longmon" változók kerültek be az elemzésbe:

Structure Matrix

	Function	
	1	2
Toll free last month	,878*	,479
Equipment last month[b]	,379*	,323
Wireless last month[b]	,089*	-,062
Long distance last month	,199	,980*
Calling card last month[b]	,210	,329*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

b. This variable not used in the analysis.

Az egydimenziós normális eloszláshoz kapcsolódó tesztstatisztika értékeket a következő menüpont választásával lehet számolni:

Analyze → Descriptive Statistics → Explore …

A menüpont választása után megjelenő ablakban a „Dependent List:" felirathoz a „tollmon" és a „longmon" változókat, a „Factor List:" felirathoz pedig a „custcat" változót helyezzük el. A „Plots …" gomb megnyomása után megjelenő ablakban ezután a „Normality plots with tests" lehetőséget választjuk, majd a „Continue" és

„OK” gomb megnyomása következik. A számolt eredmények között a Kolmogorov-Smirnov és a Shapiro-Wilk tesztstatisztika értékek is megtalálhatók.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Customer category	Statistic	df	Sig.	Statistic	df	Sig.
Toll free last month	Basic service	,184	10	,200*	,957	10	,749
	E-service	,109	10	,200*	,981	10	,970
	Plus service	,135	27	,200*	,942	27	,138
Long distance last month	Basic service	,296	10	,013	,690	10	,001
	E-service	,187	10	,200*	,914	10	,307
	Plus service	,148	27	,134	,903	27	,015

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Ha az empirikus tesztstatisztika értékeket a szignifikancia szintként gyakran alkalmazott 5 százalékhoz (0,05) hasonlítjuk, akkor nem mindegyik változó eloszlása tekinthető normális eloszlásnak mindhárom csoportban. Ennél alacsonyabb, például 0,1 százaléknál (0,001-nél) is kisebb szignifikancia szinten azonban mindhárom csoportban és mindkét változónál elfogadható lenne az egydimenziós normális eloszláshoz kapcsolódó nullhipotézis (a Kolmogorov-Smirnov és a Shapiro-Wilk teszt esetében is). A következő feladatok értelmezésénél (erre való külön utalás nélkül is) figyelembe kell venni, hogy az elemzés alkalmazási előfeltevéseinek teljesülése – a kétdimenziós normális eloszlásra vonatkozó tesztstatisztika érték számolása nélkül – mindössze meglehetősen alacsony szignifikancia-szint választása esetében nem vethető el. A gyakorlatban hasonló helyzetben (vagyis amikor nem teljesül a normális eloszlásra vonatkozó nullhipotézis 5 százalékos szignifikanciaszinten valamely változó esetében) esetenként érdemes lehet változó-transzformációt végezni a diszkriminancia-elemzés elvégzése előtt.

2. feladat:

Maximum mennyi dimenziós kanonikus tér hozható létre az 1. feladat adatai alapján?

A feladat megoldása:

A diszkriminancia-elemzés alapfeladata olyan diszkrimináló függvények meghatározása (az eredeti „magyarázó” változók lineáris kombinációiként), hogy az elemzésben szereplő csoportokat a lehető legjobban el lehessen különíteni a kanonikus térben. A kanonikus tér dimenziói a diszkrimináló függvényekhez kapcsolódnak, tehát a kanonikus tér maximális dimenziószáma a diszkrimináló függvények maximális számával egyezik meg: $\min\{g-1, p\}$. Az 1. feladat adatai alapján $\min\{g-1, p\} = \min\{3-1,2\} = 2$, vagyis maximum két dimenziós lehet az 1. feladat adatai alapján létrehozható kanonikus tér.

3. feladat:

Mennyi sajátérték számolható a $B^{-1} \cdot K$ mátrix esetében az 1. feladat adatai alapján?

A feladat megoldása:

A 2. feladathoz hasonlóan a sajátértékek száma szintén összefügg a diszkrimináló függvények számával. A $B^{-1} \cdot K$ mátrix sajátértékeinek száma $B^{-1} \cdot K$ mátrix rangjával egyezik meg, vagyis ebben a feladatban $\min\{g-1, p\} = \min\{3-1,2\} = 2$. A két sajátértéket az 1. feladatban számolt eredmények között található táblázat tartalmazza (a sajátértékeket a következő táblázatban bekarikázás jelöli):

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,461[a]	81,3	81,3	,562
2	,106[a]	18,7	100,0	,309

a. First 2 canonical discriminant functions were used in the analysis.

4. feladat:

Elméletileg lehetséges, hogy a diszkrimináló függvények maximális száma az elemzésben szereplő változók számával egyezik meg (a csoportosítást mutató eredményváltozó nélkül számítva)?

A feladat megoldása:

Igen, mivel a diszkrimináló függvények maximális száma $\min\{g-1, p\}$, ahol p jelöli az elemzésben szereplő „magyarázó" változók számát. Ha tehát $p < g-1$, akkor a diszkrimináló függvények maximális száma is p.

5. feladat:

Az 1. feladat adatai alapján van olyan csoport az elemzésben, amelynek csoportcentroidja a kanonikus tér mindkét dimenziójában pozitív koordinátákkal rendelkezik?

A feladat megoldása:

A diszkriminancia-elemzésben az egyes megfigyelések és a csoportcentroidok koordinátáit a kanonikus térben is ki lehet számolni. A csoportcentroidok kanonikus térbeli koordinátáit az 1. feladat megoldásaként számolható következő táblázat tartalmazza:

Functions at Group Centroids

Customer category	Function 1	Function 2
Basic service	-,792	-,472
E-service	-,734	,493
Plus service	,565	-,008

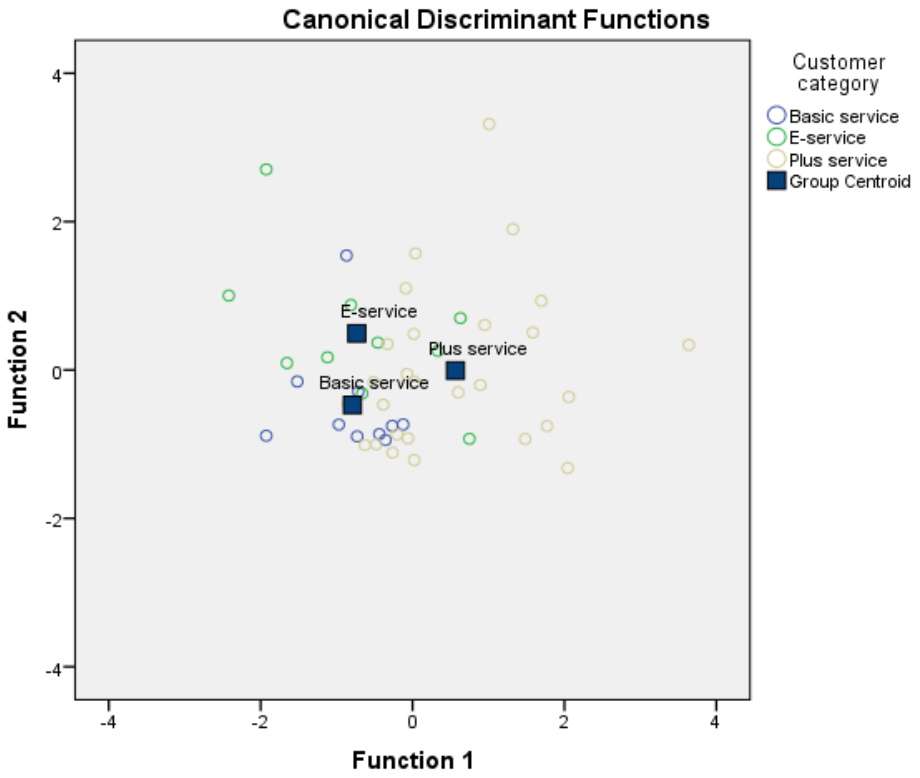
Unstandardized canonical discriminant functions evaluated at group means

Az eredmények szerint például a “Basic service” nevű csoportban a csoportcentroid koordinátái a kanonikus tér mindkét dimenziója esetében negatív előjelűek (ezeket a koordináta-értékeket a fenti táblázatban bekarikázás jelzi). A csoportok között nincs olyan, amelynek a csoportcentroidja a kanonikus tér mindkét dimenziójában pozitív koordinátákkal rendelkezik.

Az eredmények grafikusan is szemléltethetők, ha kissé kiegészítjük a diszkriminancia-elemzés 1. feladatban alkalmazott beállításait a következő menüpont választásával:

Analyze → Classify → Discriminant …

Az e menüpont választása után megjelenő ablakban a „Classify …” gombra kattintva a „Plots” feliratnál a „Combined-groups” lehetőséget választva, majd a „Continue” és az „OK” gombokat megnyomva az eredmények között megtalálható a következő ábra is, amely az elemzésben szereplő megfigyelések koordinátáit ábrázolja az 1. feladat megoldásaként számolható kétdimenziós kanonikus térben:



6. feladat:

Az 1. feladat adatai alapján mennyi a helyesen besorolt megfigyelések száma összesen?

A feladat megoldása:

A diszkriminancia-elemzésben lehetőség van az egyes megfigyelések klasszifikálására, vagyis csoportokba sorolására. Azoknál a megfigyeléseknél, amelyeknél nem ismert, hogy melyik csoportba tartoznak, a diszkriminancia-elemzés eredményei alapján egyfajta „előrejelzés” adható a csoportba tartozásra vonatkozóan,

míg a már ismert tényleges csoporttagsági adatok alapján mérni lehet a diszkriminancia-elemzés klasszifikációs teljesítményét is.

A diszkriminancia-elemzésben elméletileg többféle módon is megoldható az elemek csoportba sorolása:

- a kanonikus térbeli koordináták esetében az adott megfigyelés és a csoportcentroidok távolságának mérése után megállapítható, hogy melyik csoportcentroidtól vett távolság a minimális, és ez határozhatja meg a besorolást
- Fisher-féle diszkrimináló függvények értékei is számolhatók az egyes csoportokra külön-külön és a különböző csoportok esetén becsült értékek közül a legnagyobb érték határozhatja meg a besorolást
- minden megfigyelésnél számolhatók valószínűségek az egyes csoportokba tartozásra vonatkozóan, és ezek közül a legnagyobb érték is meghatározhatja a besorolást (ebben az esetben a „prior" valószínűségekkel kapcsolatos beállítások is befolyásolhatják a „posterior" valószínűséggel kapcsolatban számolt valószínűség-értékeket).

Az 1. feladatnál alkalmazott beállításokat kiegészíthetjük úgy, hogy a „Classify" gomb megnyomása után megjelenő ablakban a „Display" feliratnál a „Summary table" lehetőséget választjuk, majd a „Continue" és „OK" gombokra kattintunk, és így az eredmények között megtalálható a klasszifikációs eredményeket összegző következő táblázat is:

Classification Results[a]

			Predicted Group Membership			
		Customer category	Basic service	E-service	Plus service	Total
Original	Count	Basic service	9	1	0	10
		E-service	1	6	3	10
		Plus service	9	4	14	27
	%	Basic service	90,0	10,0	,0	100,0
		E-service	10,0	60,0	30,0	100,0
		Plus service	33,3	14,8	51,9	100,0

a. 61,7% of original grouped cases correctly classified.

Az összesen helyesen klasszifikált megfigyelések száma tehát (az előző táblázatban bekarikázással jelölt értékek alapján számolva): $9+6+14=29$.

7. feladat:

Hogyan értékelhető az 1. feladat eredményeinek megfelelősége a Wilks-lambda mutatószám alapján?

A feladat megoldása:

A diszkriminancia-elemzésben Wilks-lambda mutatószám számolható az egyes diszkrimináló függvényekkel kapcsolatban is, és ebben az esetben ez a Wilks-lambda mutatószám a „meg nem magyarázott heterogenitás" egyfajta mérőszáma, amelynek

kisebb (nullához közelibb) értékei az eredmények jobb, míg a magasabb (egyhez közeli) értékei az eredmények kisebb mértékű megfelelőségére utalnak.

Az 1. feladat adatai alapján a diszkrimináló függvényekhez kapcsolódó Wilks-lambda értékeket a következő táblázatban a bekarikázással jelölt értékek mutatják:

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,619	20,859	4	,000
2	,904	4,376	1	,036

Mivel az 1. feladatban legfeljebb két diszkrimináló függvényt lehetett számolni, így ebben a feladatban is két Wilks-lambda érték szerepel az eredmények között. A két szám közül az alacsonyabb értékű (0,619) a két diszkrimináló függvény által együttesen meg nem magyarázott heterogenitásra utal, mivel ennek értékét a 3. feladatban említett sajátértékek alapján a következőképpen lehet számolni (figyelembe véve hogy az eredményeket tartalmazó táblázatokban bizonyos számú tizedesjegyre kerekített értékek találhatók):

$$0{,}619 = \frac{1}{1+0{,}461} \cdot \frac{1}{1+0{,}106}$$

A két Wilks-lambda érték közül a magasabb értékű számolásához a 3. feladat megoldásában említett sajátértékek közül mindössze az egyik szükséges, mivel a magasabb értékű Wilks-lambda ebben a feladatban mindössze a második diszkrimináló függvényhez kapcsolódik:

$$0{,}904 = \frac{1}{1+0{,}106}$$

Mivel a legalacsonyabb Wilks-lambda érték is meglehetősen magas (nullához nem közeli) érték, így ebben a feladatban a klasszifikációval kapcsolatos eredmények viszonylag gyengének tekinthetők.

8. feladat:

Milyen összefüggés van a második diszkrimináló függvényhez tartozó Wilks-lambda és kanonikus korreláció értékek között az 1. feladat adatai alapján?

A feladat megoldása:

A kanonikus korreláció értékeit a diszkrimináló függvényekhez tartozó Wilks-lambda értékekhez hasonlóan a $B^{-1} \cdot K$ mátrix sajátértékei alapján lehet számolni. A $B^{-1} \cdot K$ mátrix mindegyik sajátértékéhez tartozóan külön-külön lehet számolni kanonikus korreláció értéket, amely azt mutatja, hogy a diszkrimináló „score" értékek változékonyságát milyen mértékben magyarázza a csoportbesorolás. Ezen

értelmezésből adódóan a kanonikus korreláció esetében a magas (egyhez közeli) értékek a diszkriminancia-elemzés eredményeinek nagyobb mértékű megfelelőségére utalnak. A kanonikus korreláció értékeit az eredmények között található következő táblázat tartalmazza:

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,461[a]	81,3	81,3	,562
2	,106[a]	18,7	100,0	,309

a. First 2 canonical discriminant functions were used in the analysis.

A 7. feladat megoldása alapján a második diszkrimináló függvényhez tartozó Wilks-lambda érték és az alacsonyabbik sajátérték kapcsolatát a $0{,}904 = \frac{1}{1+0{,}106}$ összefüggés írja le. Ebben a feladatban a második diszkrimináló függvényhez tartozó kanonikus korrelációs érték (0,309) és Wilks-lambda (0,904) összefüggése tehát a $B^{-1} \cdot K$ mátrix legkisebb sajátértéke (0,106) alapján számolva (figyelembe véve hogy az előző táblázatokban az eredmények csak meghatározott számú tizedesjegyre kerekítve találhatók meg):

$$0{,}309 = \sqrt{1-0{,}904} = \sqrt{1-\frac{1}{1+0{,}106}} = \sqrt{\frac{0{,}106}{1+0{,}106}}$$

Gyakorló feladatok

1. Végezzen diszkriminancia-elemzést a „longmon”, „tollmon”, „cardmon” és „wiremon” változók (magyarázó változók), valamint a „custcat” változó (függő változó) alapján stepwise (Wilks’ lambda, „Use probability of F”) módszerrel az összes adat esetében. Teljesülnek az elemzés alkalmazási előfeltevései?

2. Legfeljebb mennyi diszkriminánló függvény számolható a *Gyakorló feladatok* 1. feladatának adatai alapján? Hogyan függ össze ez az érték a „custcat” változó kategóriáinak számával?

3. Mentse el a *Gyakorló feladatok* 1. feladatának adatai alapján számolható diszkriminánló „score”-ok értékeit. Mennyi az elmentett új változók esetében az átlag értéke?

4. Elméletileg befolyásolja a $B^{-1} \cdot K$ mátrix legnagyobb sajátértékének értéke a diszkriminancia-elemzési modellben az összes diszkriminánló függvény alkalmazásával számolható Wilks-lambda értéket?

5. Elméletileg legfeljebb mennyi Fisher-féle diszkriminánló függvény számolható valamely diszkriminancia-elemzésben?

6. Ha egy diszkriminancia-elemzésben több mint egy dimenziós a kanonikus tér, akkor elméletileg mennyi a kanonikus tér egyes dimenzióihoz tartozó diszkrimináló „score”-ok közötti lineáris korrelációs együttható értéke?

Irodalomjegyzék

Hajdu Ottó [2003]: *Többváltozós statisztikai számítások*
Központi Statisztikai Hivatal

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Mardia, K.V. [1970]: *Measures of Multivariate Skewness and Kurtosis with Applications.* Biometrika, Vol.57., No. 3. pp. 519-530.

McNeil, A.J. - Frey, R. - Embrechts, P. [2005]: *Quantitative Risk Management: Concepts, Techniques and Tools.*
Princeton University Press

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. A diszkriminancia-elemzés alkalmazási előfeltevése

   a) a csoportokban a változók varianciái megegyeznek
   b) a csoportokban a változók kovarianciái megegyeznek
   c) mindkét előző állítás helyes
   d) egyik előző állítás sem helyes.

2. A diszkriminancia-elemzésben a maximálisan képezhető diszkrimináló függvények száma (ha *p* a változók számát, *g* pedig a csoportok számát jelöli)

   a) min(p,g-1)
   b) min(p-1,g)
   c) min(p,g)
   d) min(p-1,g-1)

3. A diszkrimináló függvények együttes szétválasztó „ereje” jónak tekinthető, ha a Wilks-lambda értéke

   a) 1-hez közeli érték
   b) nullához közeli érték
   c) -1-hez közeli érték
   d) egyik előző válasz sem helyes.

4. A diszkriminancia-elemzésben a struktúra mátrixban

   a) pozitív és negatív értékek és nulla is szerepelhetnek
   b) nulla kivételével mindenféle érték szerepelhet
   c) csak pozitív értékek szerepelhetnek
   d) csak negatív értékek szerepelhetnek

5. Ha legfeljebb egyetlen diszkrimináló függvény képezhető, az ehhez a diszkrimináló függvényhez tartozó Wilks-lambda

   a) értéke alapján kiszámolható a kanonikus korreláció értéke is
   b) lehet negatív érték is
   c) nem lehet egynél kisebb érték
   d) egyik előző válasz sem helyes.

8. fejezet

SOKDIMENZIÓS SKÁLÁZÁS

A módszer rövid összefoglalása

A többdimenziós skálázás sokoldalúan alkalmazható adatelemzési módszer. A sokféle skálázási modell közül ebben a fejezetben az ALSCAL és az INDSCAL módszerekkel foglalkozunk. Az ALSCAL skálázási módszernél – az eredeti különbözőségek és a származtatott koordináták közötti távolságok eltéréseinek minimalizálásával – az adatok között mért különbözőségek alapján származtatunk koordinátákat a skálatérképen. Ez az elemzés bizonyos szempontból a főkomponens-elemzéshez, egy másik szempontból pedig a klaszterelemzéshez hasonló, mivel a többdimenziós skálázással feltárhatók a változók és a megfigyelések közötti egyes összefüggések is. Az ALSCAL skálázási modellel szemben az INDSCAL skálázásnál különböző csoportokhoz külön-külön távolságmátrixot lehet számolni, és az eredmények alapján (a gyakorlati modellbeállításoknak megfelelően) a változók, illetve a megfigyelések kapcsolatrendszerének csoportok közötti eltéréseivel összefüggő következtetések adódhatnak. Az INDSCAL skálázásnál a csoportok esetleges különbözőségei például az egyedi terek és a csoport tér összehasonlításával mutathatók be.

A többdimenziós skálázásos modellt nem-metrikusnak nevezzük, ha a skálatérképen a távolságok ordinálisan kapcsolódnak az eredeti különbözőségekhez. Metrikus skálázás esetén a skálatérképen a távolságok és az eredeti különbözőségek között lineáris függvénykapcsolat van, ebben az esetben a modell intervallum vagy arány skálájú lehet. A metrikus és az ordinális skálázás hasonló eredményre vezet, ha euklideszi távolságokból indulunk ki, nem euklideszi távolságnál ugyanakkor a nem-metrikus skálázás alkalmazása javasolható. Az illeszkedés jóságát a STRESS (standardized residual sum of squares) függvény értéke méri.

Megoldási módszerek és az eredmények értelmezése

Gyakorlati szempontból előnyös lehet, hogy többdimenziós skálázást többféle mérési szintű változóval is lehet végezni, a feladatok megoldásakor a modellbeállítások során ugyanakkor a változók mérési szintjét is figyelembe kell venni (például a „távolság” mérésekor). A többdimenziós skálázás során alkalmazott változókat gyakran szokás valamilyen módon transzformálni, például sztenderdizálni. A következő feladatokban néhány magas (arány) mérési szintű változó alapján végzett elemzéssel a többdimenziós skálázás egyes alkalmazási lehetőségeit szemléltetjük.

Az ALSCAL modellnél nincs szerepe az adatok csoportokba sorolásának, az INDSCAL modell alapján készített elemzéseknél azonban a csoportok közötti különbségeknek kiemelt jelentősége van. Az ALSCAL és INDSCAL modellekkel számolt eredmények egyszerűbb összehasonlíthatósága érdekében a következő feladatok során a teljes adathalmaz egy részhalmazába tartozó adatokkal foglalkozunk, (amelyek az egyik „kategóriás” változó két kiválasztott csoportjába tartoznak).

A gyakorló feladatok megoldásánál említett változók a bankloan.sav adatai között találhatók.

1. feladat:
Szűrje ki azokat az adatokat a teljes adathalmazból, amelyek esetében az „ed” változó értéke 1 vagy 2! Mennyi a szűrés után az elemzésben szereplő megfigyelések száma?

A feladat megoldása:

Az adatok szűrése a következő menüpont választásával kezdhető el:

Data → Select Cases …

A menüpont kiválasztása után megjelenő ablakban a „Select” feliratnál az „If condition is satisfied” feliratnál található „If…” gombra kattintva megjelenik egy újabb ablak, amelyben található egy – a szűrési feltételek definiálásához kapcsolódó – képletek beírására alkalmas rész, amelybe az „ed” nevű változóra hivatkozva a következő képlet írható be:

ed < 3

Az „ed<3” feltétel alapján (mivel az „ed” változó kategóriáit pozitív egész számok jelölik az adathalmazban) a feladatleírásnak megfelelő adatok szűrésére kerülhet sor.

A szűréssel kapcsolatos feladat megoldását az „ed” változóra vonatkozó (a korábbi fejezetekben leírt módon számolható) gyakorisági tábla is szemlélteti:

Level of education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Did not complete high school	460	66,2	66,2	66,2
	High school degree	235	33,8	33,8	100,0
	Total	695	100,0	100,0	

E gyakorisági tábla alapján tehát a szűrés után az elemzésben összesen 695 megfigyelés szerepel.

2. feladat:
A (sztenderdizált) „age”, „income”, „debtinc”, „creddebt” és „othdebt” változók alapján végezzen többdimenziós skálázást (ALSCAL módszerrel, ahol a változókra ordinális modellt alkalmaz úgy, hogy a dimenziószám 2, a modellben a távolságot pedig euklideszi távolsággal méri)! Milyen a modell illeszkedése a STRESS mutatószám alapján?

A feladat megoldása:

A többdimenziós skálázási feladatok megoldása során számos modellbeállítással kell foglalkozni. Az egyik legfontosabb paraméter a megoldás során a dimenziószám (ez ebben az esetben 2), amely arra utal, hogy a feladatban szereplő 5 változó alapján mennyi dimenziós skálatérképre vonatkozóan szeretnénk koordinátákat származtatni (ebben a feladatban a skálatérkép 2 dimenziós).

A többdimenziós skálázás például olyan szempontból is hasonlít a hierarchikus klaszterelemzésre, hogy az elemzésben nagy szerepe van bizonyos távolságmátrixoknak, amelyek számításához többféle távolságot mérő mutatószám is alkalmazható. Az euklideszi távolságon kívül tehát a feladatban szereplő (arány mérési szintű) változók esetében más mutatószámot is lehetne választani, az euklideszi távolság mindössze az egyik lehetőség.

Az ALSCAL modell elméletileg ordinális, intervallum és arány skálájú lehet, és e három lehetőség között az a különbség, hogy milyen összefüggés van a modellben a skálatérképen mért távolságok és az eredeti különbözőségek között. Az ordinális modell esetén ebben a feladatban a skálatérképen a távolságok és az eredeti különbözőségek között olyan összefüggés kialakítása a cél, hogy ha az eredeti különbözőség-értékek közül két kiválasztott értéknél például az első nem nagyobb mint a második, akkor a skálatérképen számított távolságok esetében se legyen nagyobb a két megfelelő (az eredeti különbözőség-értékekhez tartozó) érték közül az első.

Az ALSCAL modellhez szükséges beállítási lehetőségeket a következő menüpont kiválasztásával lehet megtekinteni:

Analyze → Scale → Multidimensional Scaling (ALSCAL) …

Az ezt követően megjelenő ablakban a „Variables:” felirat alatt elhelyezhető a feladatban szereplő öt változó. Mivel a feladatban nem eleve távolságmátrix szerepel, így azt először szükséges definiálni a feladat megoldása során, amely megoldható az ennek az ablaknak az alján található „Create distance from data” lehetőségnél a „Measure …” gombra való kattintással, majd ezek után a „Measure” felirat alatt az „Interval:” feliratnál az „Euclidean distance” lehetőség kiválasztásával.

Mivel a feladatban sztenderdizált változókra van szükség, így a távolságmátrix definiálását követően az aktuális ablakban a „Transform Values” feliratnál a „Standardize:” feliratnál a „Z scores” lehetőség választása következik (a „By variable” lehetőség választásával). A többdimenziós skálázás során elméletileg a változók és a megfigyelések esetében is lehetne távolságmátrixot létrehozni. E két lehetőség közül ebben a feladatban a változókkal foglalkozunk, így a „Create Distance Matrix” feliratnál a „Between variables” lehetőség választása szükséges. E beállítások után a „Continue” gombra lehet kattintani.

A skálatérkép dimenziószáma és néhány egyéb paraméter a „Model” gombra kattintás után határozható meg. Ebben a feladatban a megoldások az ezután megjelenő ablakban a különböző feliratoknál a következő lehetőségek választásával számolhatók:

- „Level of Measurement” feliratnál: „Ordinal” lehetőség választása (a feladatleírásnak megfelelően)
- „Dimensions” feliratnál: a „Minimum:” és a „Maximum:” érték egyaránt 2, így a feladatmegoldás során mindössze a két dimenzió esetére vonatkozó eredmények számolódnak (elméletileg nem szükséges hogy e két szám megegyezzen, a minimum érték kisebb is lehet, mint a maximum érték, és ekkor – ha egyéb

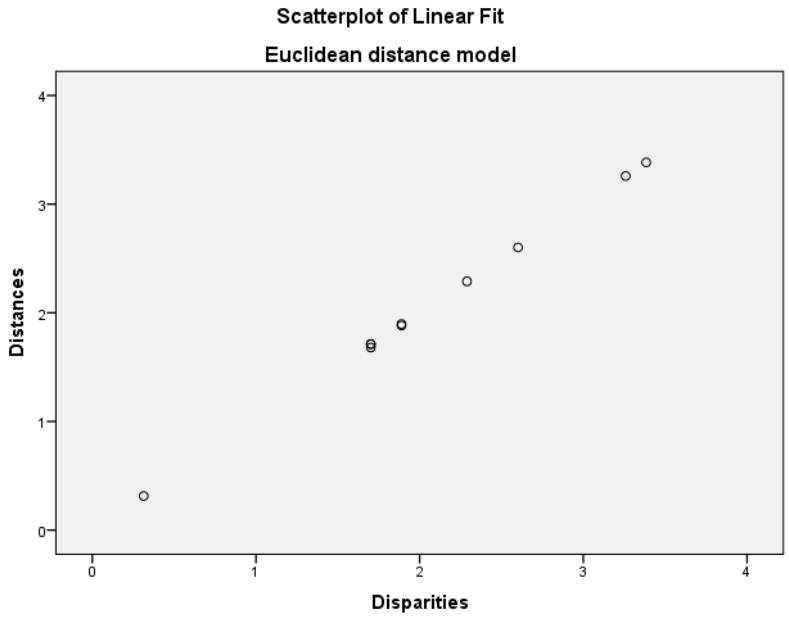
feltételeknek is megfelelnek a modellbeállítások – egyszerre több dimenzióra vonatkozó eredmények is számolhatók)
- „Scaling Model:” feliratnál: az „Euclidean distance” lehetőség választása
- „Conditionality:” feliratnál: „Matrix” lehetőség választása (előfordulhatna, hogy a távolságmátrix elemeinek értelmezése attól is függ, hogy az egyes elemek a távolságmátrix melyik részében találhatók, az alapértelmezésnek is megfelelő „Matrix” lehetőség választásakor azonban szimmetrikusnak feltételezzük a távolságmátrixot).

E beállítások elvégzése után a „Continue” gombra lehet kattintani. Az eredmények grafikus megjelenítése megkönnyítheti az eredmények értelmezését, így az „Options:” gombra kattintva a „Display:” feliratnál a „Model and options summary” lehetőségen kívül a „Group plots” lehetőséget is kiválaszthatjuk. Ezután a „Continue” gombra, majd az „OK” gombra való kattintás után megtekinthetők a feladatmegoldáshoz kapcsolódó eredmények.

A modell illeszkedésének megfelelőségét a STRESS mutatószám alapján lehet értékelni, a következő modell-output esetében a bekarikázott érték mutatja, hogy a feladatban az alkalmazott paraméterek esetében a STRESS mutató értéke 0,0039:

```
                For matrix
Stress  =    ,00390        RSQ =  ,99989


          Configuration derived in 2 dimensions


                  Stimulus Coordinates

                       Dimension

Stimulus   Stimulus      1         2
 Number      Name

    1       age         1,4277   -1,0982
    2       income      1,2945     ,7933
    3       debtinc    -1,7919    -,5943
    4       creddebt    -,5738     ,5631
    5       othdebt     -,3565     ,3360
```

A STRESS mutató 0,05 alatti értéke kiváló illeszkedésre utal, vagyis ekkor a skálatérképen a távolságok és az eredeti különbözőségek között nagymértékű a kapcsolat. A feladat eredményei között ugyanerre a jelenségre (a skálatérképen mért távolságoknak az eredeti különbözőségekhez való kiváló illeszkedésére) utal a lineáris illeszkedéshez kapcsolódó R-négyzet viszonylag magas (0,99989) értéke is. Az eredeti különbözőségek függvényében a skálatérképen mért távolságok értékeit mutatja a következő ábra:



Az eredeti különbözőségek és a skálatérképen mért távolságok illeszkedését mutató grafikonon viszonylag kevés ábrázolt pont található, ami azzal függ össze, hogy mindössze öt változó szerepelt az elemzésben, és ebben a feladatban a változókra vonatkozóan került sor a többdimenziós skálázás eredményeinek számolására (így például a skálatérképen mért távolságok értékeit tartalmazó távolságmátrixnak 5 sora és 5 oszlopa van).

3. feladat:

A 2. feladat megoldása alapján melyik két változó tekinthető az egymáshoz leginkább hasonlónak?

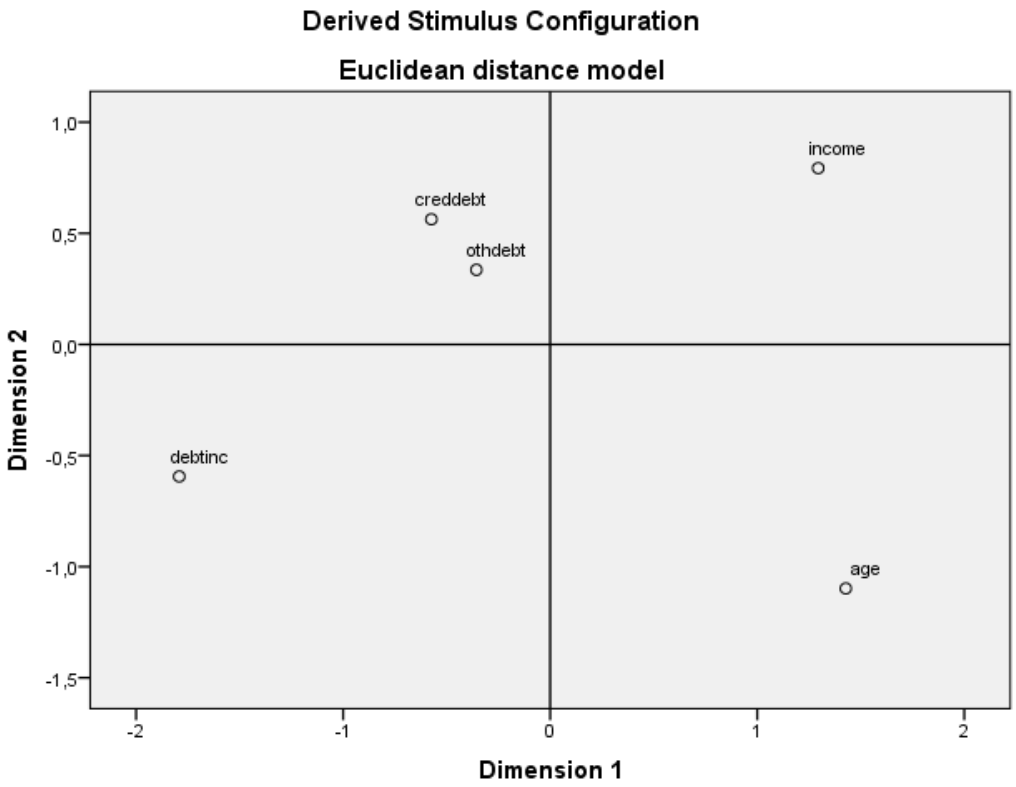
A feladat megoldása:

A feladatban az a két változó tekinthető az egymáshoz leginkább hasonlónak, amelyek a kétdimenziós skálatérképen a legközelebb vannak egymáshoz. A kétdimenziós eredményeknél a „creddebt" és az „othdebt" változók (bekarikázással jelölt) koordinátái között kisebb az eltérés, mint más változópárok koordinátái esetében, így ez a két változó tekinthető egymáshoz leginkább hasonlónak:

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	2
1	age	1,4277	-1,0982
2	income	1,2945	,7933
3	debtinc	-1,7919	-,5943
4	creddebt	-,5738	,5631
5	othdebt	-,3565	,3360

Ez az eredmény szemléltethető a skálatérképen is (ebben az esetben a kétdimenziós koordináták alapján ábrázolva a változókat):



4. feladat:

A 2. feladat adatai alapján végezze el az elemzést arány modell esetében is (amikor az ALSCAL modell arány skálájú). Hasonlítsa össze a STRESS mutató értéket a 2. feladat eredményével és értelmezze az esetleges különbséget!

A feladat megoldása:

Az arány mérési szint beállítása arra utal, hogy elméletileg ekkor az eredeti különbözőségek és a skálatérképen mért távolságok között olyan lineáris kapcsolatot feltételezünk, amelynél a lineáris összefüggést leíró függvényben a konstans nulla. Ebben az esetben az ordinális modellhez viszonyítva kissé nehezebb lehet olyan skálatérképen mért távolságokat meghatározni, amelyek kiválóan illeszkednek az eredeti különbözőségek értékeihez.

Ebben a feladatban a 2. feladatban alkalmazott beállítások alkalmazhatók azzal a változtatással, hogy a „Model” gombra kattintás után a „Level of Measurement” feliratnál az „Ordinal” lehetőség helyett a „Ratio” lehetőség választásával állítható be az ALSCAL modell arány skálája.

A 2. feladat megoldásához hasonlóan a modell illeszkedésének megfelelőségét a STRESS mutatószámmal lehet mérni. Ha a STRESS mutatószám értéke 0,05 alatti, akkor jónak tekinthető az illeszkedés (vagyis a skálatérképen mért távolságok és az eredeti különbözőségek között meglehetősen szoros a kapcsolat). A STRESS mutató

magasabb értéke gyengébb illeszkedésre utal. Ebben a feladatban a modell illeszkedése gyengébbnek tekinthető, mint a 2. feladatban, mivel a STRESS mutató értéke 0,20029:

```
                For  matrix
Stress  =     ,20029        RSQ =  ,81059
```

```
        Configuration derived in 2 dimensions
```

```
                Stimulus Coordinates

                     Dimension

Stimulus   Stimulus     1        2
 Number      Name

    1       age        1,3137  -1,2074
    2       income     1,2957    ,7605
    3       debtinc   -1,5688   -,8208
    4       creddebt   -,6316    ,7991
    5       othdebt    -,4089    ,4686
```

Az eredmények azt is mutatják, hogy a kétdimenziós skálatérképen a változók koordinátái is eltérőek az arány és az ordinális modell esetében, ugyanakkor az ordinális modellhez hasonlóan az arány skálájú modellnél is a "creddebt" és az "othdebt" változók tekinthetők egymáshoz a leginkább hasonlónak.

5. feladat:

A 2. feladat adatai alapján végezze el az elemzést (két dimenziós modell helyett) egy dimenziós modell esetében is. Hasonlítsa össze a STRESS mutató értéket a 2. feladat eredményével és értelmezze az esetleges különbséget!

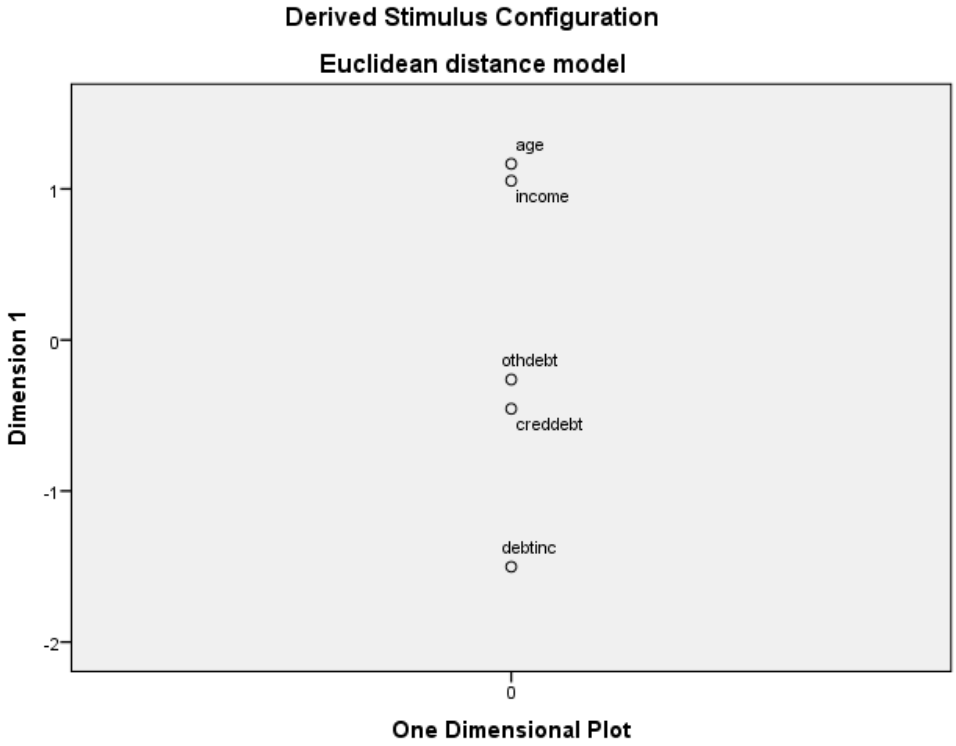
A feladat megoldása:

A feladatmegoldás során a 2. feladatban alkalmazott beállítások változatlanok maradhatnak, kivéve hogy a „Model" gombra kattintás után a „Dimensions" feliratnál a „Minimum:" és a „Maximum:" érték egyaránt 1 értékű.

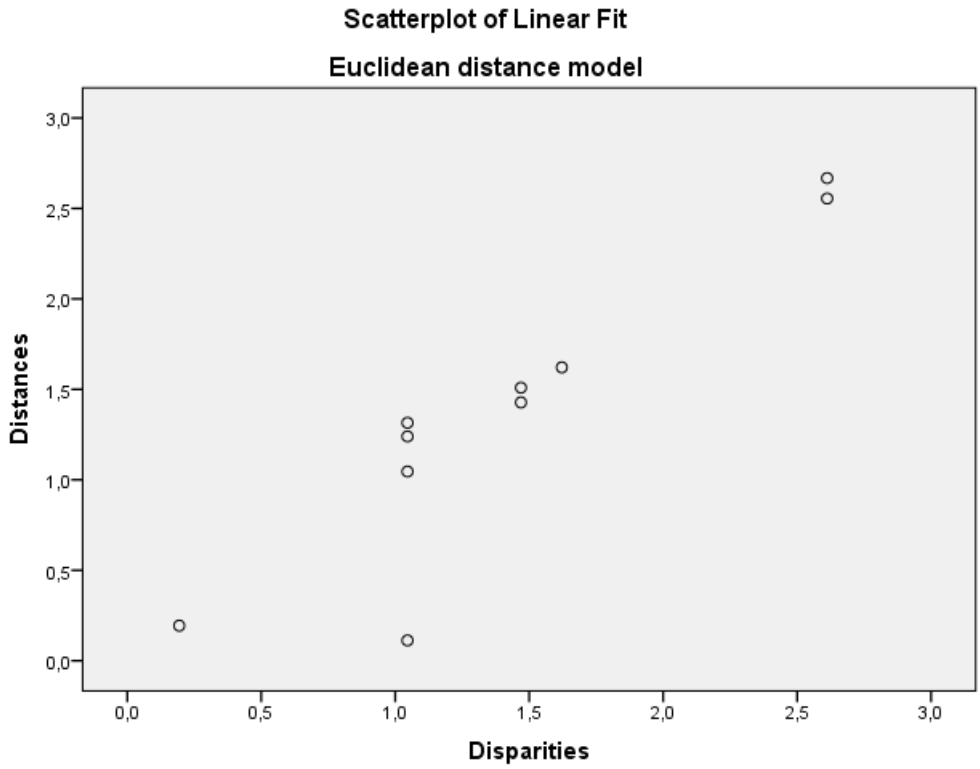
Az eredmények között található STRESS mutató értéke az 1 dimenziós modellnél 0,19841 (ezt a következő értékek között található bekarikázott érték is mutatja). Ez az érték nagyobb, mint a 2. feladat megoldásánál számolt STRESS mutatószám értéke, tehát az egy dimenziós modell illeszkedése gyengébb a 2 dimenziós modellnél ebben a feladatban.

```
                For  matrix
Stress  =    ,19841        RSQ =  ,84631


          Configuration derived in 1 dimensions


                  Stimulus Coordinates

                  Dimension

Stimulus   Stimulus     1
 Number      Name

    1       age        1,1659
    2       income     1,0534
    3       debtinc   -1,5016
    4       creddebt   -,4558
    5       othdebt    -,2619
```

Az egy dimenziós skálatérképen a változók elhelyezkedése is különbözik a 2 dimenziós skálatérképtől, az egyik különbség például, hogy nem a „creddebt” és az „othdebt” változók tekinthetőek az egymáshoz leginkább hasonlónak (mivel nem e két változó távolsága a legkisebb, ahogyan arra az egydimenziós koordináták alapján is következtetni lehet).



Az eredeti különbözőségek függvényében az egy dimenziós skálatérképen mért távolságok értékeit mutatja a következő ábra:



Az eredeti különbözőségek és az egy dimenziós skálatérképen mért távolságok közötti összefüggés tehát kevésbé szoros, mint a 2. feladatban, ahogyan erre a lineáris illeszkedéshez kapcsolódó R-négyzet értéke is utal (az R-négyzet mutató értéke az egydimenziós modellben 0,84631).

6. feladat:

A (sztenderdizált) „age", „income", „debtinc", „creddebt" és az „othdebt" változók alapján az „ed" változó két kategóriájának megkülönböztetésével végezzen többdimenziós skálázást INDSCAL módszerrel (változók elemzése kétdimenziós ordinális modellel)! A két dimenzió közül melyik a fontosabb?

A feladat megoldása:

Az INDSCAL modell esetében az ALSCAL modellhez hasonlóan meghatározott értéknek megfelelő dimenziószámú térbeli koordinátákat lehet számolni, azonban az ALSCAL modellel ellentétben az INDSCAL modellben az elemzésben szereplő minden csoportra külön-külön térbeli koordináták számolhatók. Ezeken az „egyedi tereken" kívül az INDSCAL modellben számolhatók még „csoport térbeli" koordináták úgy, hogy az egyedi terek és a csoporttér közötti kapcsolatokat az úgynevezett egyedi súlyok írják le. Az INDSCAL modellben tehát a csoportosítást

leíró változónak kiemelt szerepe van, így az INDSCAL modellnél az elemzésben szereplő egyik változó a csoportosítást írja le (kategóriái az elemzésben szereplő különböző csoportokhoz tartoznak).

Az INDSCAL modellhez tartozó beállítási lehetőségeket az ALSCAL modellhez hasonlóan a következő menüpont választásával lehet megtekinteni:

Analyze → Scale → Multidimensional Scaling (ALSCAL) …

A feladatmegoldás során a 2. feladatban alkalmazott beállítások változatlanok maradhatnak, két változtatástól eltekintve:

- az előzőekben jelzett menüpont választása után megjelenő ablakban az „Individual Matrices for:" felirat után ebben a feladatban az „ed" változó kerül
- a „Model" gombra kattintás után a „Scaling Model" feliratnál az „Individual differences Euclidean distance:" lehetőség választására kerül sor úgy, hogy ebben a feladatban az „Allow negative subject weights" lehetőség nincs bejelölve.

Az eredmények között találhatók a két dimenzió „fontosságát" mérő értékek:

Subject Weights

Subject Number	Weirdness	Dimension 1	Dimension 2
1	,3180	,9086	,4170
2	,5458	,9921	,1007
Overall importance of each dimension:		,9049	,0920

Ebben a feladatban az első dimenzió tekinthető a fontosabbnak, mivel az eredmények között bekarikázott két érték esetében $0{,}9049 > 0{,}0920$.

7. feladat:

A 6. feladat adatai alapján INDSCAL modell alkalmazásával végzett többdimenziós skálázásnál hogyan állapítható meg, hogy a két csoporthoz tartozó súlyok az átlagos súlyokkal arányosak-e?

A feladat megoldása:

Az INDSCAL modellben az egyedi súlyok az egyedi terek és a csoport tér közötti kapcsolatot írják le. Valamely súly az adott csoportra és a közös dimenziós térben számított MDS koordinátákra vonatkozik, a súlyok értéke 0 és 1 közötti lehet. Az INDSCAL modellben számolható súlyoknál azonban nem elsősorban az egyes súlyok értékei, hanem (a „súly-térben") az egyedi súlyok közötti szögek értelmezhetők.

A súlyok terénél értelmezhető például az origóból a súlyt jelölő ponthoz húzott vektorok között bezárt szög. Ha ez a bezárt szög kicsi két súly-vektor között, akkor a két csoportban hasonlóan súlyozódnak a dimenziók. Az INDSCAL modellben szintén a súly-vektorok alapján számolható a „weirdness" index, amelynek értéke elméletileg 0 és 1 közötti lehet úgy, hogy a nulla érték utal arra az esetre, amikor valamely elemzésben szereplő csoport súlyai az átlagos súlyokkal arányosak.

A 6. feladat megoldásában szereplő eredmények közül a „Weirdness" értékek közül egyik sem nulla, így egyik csoport súlyai sem arányosak az átlagos súlyokkal.

Subject Weights

Subject Number	Weird-ness	Dimension 1	Dimension 2
1	,3180	,9086	,4170
2	,5458	,9921	,1007
Overall importance of each dimension:		,9049	,0920

8. feladat:

A 6. feladat adatai alapján INDSCAL modellel végzett többdimenziós skálázásnál mennyi az egyes dimenziókhoz rendelhető lapított súlyok összege?

A feladat megoldása:

Az INDSCAL elemzésben a súlyok terében az origóból a súlyokat jelölő pontokhoz húzott vektorok között bezárt szögek alapján lehet számolni az úgynevezett „lapított súlyokat", amelyeket a feladat eredeti dimenziószámához képest eggyel kevesebb dimenziószámú térben lehet ábrázolni. Ebben a feladatban grafikusan úgy lehetne ábrázolni a lapított súlyok számolását, hogy a feladatban szereplő két dimenziós súly-térben a mindkét tengellyel 45°-os szöget bezáró egyenesre lehetne vetíteni az eredeti súlyokat jelölő pontokat. A lapított súlyok összege minden tengely esetében nulla. Ezt az összefüggést mutatják a 6. feladat adatai alapján számolt következő eredmények is (a lapított súlyok értékeit bekarikázás jelöli):

Flattened Subject Weights

Subject Number	Plot Symbol	Variable 1
1	1	-1,0000
2	2	1,0000

Gyakorló feladatok

1. A (sztenderdizált) „age”, „years with current employer”, „income”, „debtinc”, „creddebt” és az „othdebt” változók alapján végezzen többdimenziós skálázást ALSCAL modellel 1, 2 és 3 dimenziós esetben (a változókra ordinális modellt alkalmazzon úgy, hogy a távolságot euklideszi távolsággal méri)! Melyik modell illeszkedése tekinthető a legjobbnak?

2. Hogyan határozható meg a *Gyakorló feladatok* 1. feladatában a leginkább megfelelő dimenziószám?

3. A *Gyakorló feladatok* 1. feladatának adatai alapján számolt eredményeknél melyik két változó tekinthető az egymástól leginkább különbözőnek?

4. Milyen összefüggés van az INDSCAL modellben az egyedi terek és a csoport tér koordinátái között?

5. Nem egyenlő elemszámú csoportoknál is lehetőség van INDSCAL modell alkalmazására a megfigyelések elemzésénél?

6. Az INDSCAL modellben elméletileg kiszámolható az egyes dimenziók fontossága az egyedi terekhez tartozó súlyok értékei alapján?

7. Az INDSCAL modellben 2 dimenziós csoport térben számított koordináták megegyeznek a 2 dimenziós ALSCAL modellben számolt skálatérképen szereplő koordinátákkal?

8. Hogyan mérhető az eredmények megfelelősége az INDSCAL elemzésben?

Irodalomjegyzék

Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése*
Tanszék Kft., Budapest

Kovács Erzsébet [2014]: *Többváltozós adatelemzés*
Typotex Kiadó, Budapest

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. Az ALSCAL elemzésben a STRESS mutatószám

   a) az R-négyzet mutatószámmal egyenlő
   b) az R-négyzet mutatószám négyzetgyöke
   c) az R-négyzet mutatószám lineáris függvénye
   d) egyik előző állítás sem helyes.

2. Ha az elemzésben szereplő változók mérési szintje megegyezik, akkor az ALSCAL elemzés milyen mérési szintű változók esetében végezhető?

   a) nominális
   b) ordinális
   c) arány
   d) mindegyik előző válasz helyes.

3. Az INDSCAL elemzés nem végezhető el, ha

   a) a csoportok száma az elemzésben egy
   b) az elemzésben szereplő változók átlaga nulla
   c) az elemzésben szereplő változók szórása egy
   d) egyik előző válasz sem helyes

4. Az ALSCAL elemzésben kétdimenziós eredményeknél a tengelyek közötti lineáris korreláció értéke lehet

   a) egy
   b) mínusz egy
   c) nulla
   d) mindegyik előző válasz helyes.

5. Az INDSCAL elemzésben bármely adathalmazban minden esetben

   a) a súlyok összege dimenziónként egységnyi
   b) a súlyok négyzetösszege dimenziónként egységnyi
   c) a súlyok négyzetösszege dimenziónként R-négyzet mutatószámként értelmezhető
   d) egyik előző válasz sem helyes.

9. fejezet

TÚLÉLÉSI MODELLEK

A modellek rövid összefoglalása

A túlélési (survival) modellek egy legfeljebb egy alkalommal bekövetkező esemény megtörténtéig eltelt idő vizsgálatára használhatók. Néhány tipikus gyakorlati példa:

1. Adott gép / eszköz élettartama (a beüzemeléstől az elromlásig eltelt idő hossza)
2. Betegség diagnosztizálásától az elhalálozásig eltelt idő hossza (tegyük fel, hogy olyan súlyos betegségről van szó, aminek következtében minden beteg meghal)
3. Betegség diagnosztizálásától a meggyógyulásig eltelt idő hossza (tegyük fel, hogy olyan betegségről van szó, aminek következtében nem hal meg senki)
4. Biztosítási szerződés kezdetétől a megszűnéséig eltelt idő hossza

A módszerek bemutatásánál az egységesség kedvéért *Vékás Péter [2011]: Túlélési modellek*[11] c. tanulmányában használt jelölésrendszert alkalmazzuk (pár kivételtől eltekintve), illetve a jobb megértés kedvéért megismétlünk néhány itt szereplő összefüggést.

Survival modellek alkalmazásához két változót szükséges definiálni: egy s státusz-, és egy t időváltozót. Amennyiben egy adott i megfigyelés esetén bekövetkezik a modellezni kívánt esemény, akkor s_i értéke 1, t_i pedig a megtörténésig eltelt idő hossza. Elképzelhető olyan eset is, amikor a megfigyelt időszak alatt nem következik be a kérdéses esemény. Az ilyen eseteket cenzorált megfigyeléseknek nevezzük, és ilyenkor s_i értéke 0, t_i pedig a megfigyelési időszak hossza. Jelölje T nemnegatív valószínűségi változó a vizsgált esemény bekövetkezéséig eltelt időt. A feladat a $G(t) = P(T \geq t)$ túlélésfüggvény becslése. Előbbi függvény minden $t \geq 0$ szám esetén annak a valószínűségét adja meg, hogy a vizsgált esemény legalább t idő elteltével következik be.

A helyes alkalmazáshoz az első lépés mindig a kérdéses esemény és a fenti változók definiálása kell hogy legyen. A módszer nevében ugyan a „túlélés” szó szerepel, ami azt sugallja, hogy a modellezett történés az elhalálozás / elromlás / megszűnés, de ez nem minden példában törvényszerű (pl. a fenti 3. példában a kérdéses esemény a meggyógyulás, és a „túlélés” pedig az, hogy továbbra is beteg marad az egyén).

1. feladat:

Adjuk meg a fenti 1. példában a modellezendő eseményt, a túlélés és cenzorált megfigyelés jelentését, a T, s_i és t_i definícióját!

A feladat megoldása:

A modellezendő történés a gép elromlása kell legyen, T jelöli a beüzemeléstől tönkremenésig eltelt időt. Ha a megfigyelt időszak alatt elromlik az eszköz, akkor s_i értéke 1, t_i pedig a gép

[11] A tanulmány a *Pénzügyi adatok statisztikai elemzése (Kovács Erzsébet [2011]) c.* egyetemi tankönyv 9. fejezetében található

élettartama. Ha pedig a gép nem megy tönkre az időszak alatt, akkor a megfigyelésünk cenzorált, $s_i = 0$, t_i pedig a megfigyelési időtartam.

A továbbiakban két túlélési modellel foglalkozunk majd: a Kaplan-Meier becsléssel és a Cox regresszióval. Előbbit akkor használjuk, ha a teljes mintára vagy annak valamilyen almintáira szeretnénk becsülni a túlélést, úgy hogy a túlélési valószínűségeket csak a státusz- és az időváltozó felhasználásával állítjuk elő, egyéb magyarázó változót nem vonunk be a modellbe. A $P(T \geq t)$ túlélésfüggvény becslése ekkor egyszerűen elkészíthető a **2.** feladat **a,** és **b,** pontjában bemutatott módszer segítségével.

Előbbivel ellentétben a Cox-regressziót akkor használjuk, ha azt tételezzük fel, hogy a $P(T \geq t)$ túlélési valószínűségeket befolyásolják az $x_1, x_2, \dots, x_p$ magyarázó változók értékei. A módszer megértéséhez szükségünk lesz néhány új fogalomra és összefüggésre. Jelölje továbbra is $G(t) = P(T \geq t)$ a T bekövetkezési időhöz tartozó túlélésfüggvényt, és legyen $F(t) = P(T < t)$ a T eloszlásfüggvénye. Vezessük be a következő jelöléseket:

- $f(t) = F'(t)$, T sűrűségfüggvénye.

- $h(t) = \frac{f(t)}{G(t)} (G(t) \neq 0)$, az úgynevezett kockázati ráta. Azt szemlélteti, hogy mennyire valószínű, hogy a modellezett esemény a t időponthoz képest „egy nagyon rövid időn" belül bekövetkezik[12], feltéve, hogy t-ig nem következett be[13]. A modell a $h(t)$-re ad majd regressziós becslést.

- $H(t) = \int_0^t h(s)ds$, az úgynevezett kumulált kockázati ráta. Számunkra annyiban lesz fontos, hogy $H(t)$ segítségével adható összefüggés a Cox-regresszió által becsült $h(t)$ és a túlélésfüggvény ($G(t)$) közt. Megmutatható hogy:

  $G(t) = e^{-H(t)}$.

A Cox-regressziós modell által becsült egyenlet:

$$h(t) = h_0(t) \cdot e^{b_1x_1 + b_2x_2 + \cdots + b_px_p}.$$

A $h_0(t)$ az úgynevezett alap kockázati ráta, az összes magyarázó változó 0 értéke esetén adja a becsült kockázati rátát. Világos az is, hogy tetszőleges $x_1, x_2, \dots, x_p$ értékek esetén a fenti képletből kiszámolható $h(t)$ becslése (és a túlélésfüggvény is). Fontos, hogy az $x_1, x_2, \dots, x_p$ értékeinek behelyettesítése után nem egy számot, hanem egy t-től függő függvényt kapunk, ami a túlélési idő eloszlásának transzformáltja. A modell mögött egy arányossági feltevés húzódik: a magyarázó változóktól nem függ a kockázati ráta alakja, azok változása csak konstansszorosára növeli / csökkenti a függvényt.
A becslési eljárás ismertetése meghaladná jelen példatár kereteit, a továbbiakban az SPSS által számolt eredményeket mutatjuk be, és értelmezzük majd.

---

[12] Matematikai formulával megfogalmazva: $P(T < t + \varepsilon \mid T \geq t) \approx \varepsilon \cdot h(t)$, ha ε elég kicsi

[13] Vegyük észre, hogy $h(t)$ lényegében ugyanazt szemlélteti, mint a Kapan-Meier becslésnél a későbbi **2, a.** feladatnál bevezetett q_t mennyiség. A különbség, hogy a Kaplan-Meier féle becslés a megfigyelt véges sok kilépési pont alapján becsüli a túlélésfüggvényt (diszkrét modell), a Cox-regresszió pedig folytonos eloszlású T valószínűségi változót feltételez, és a sűrűségfüggvényből előálló kockázati rátát becsüli.

A Kaplan-Meier modell alkalmazása és az eredmények értelmezése

A gyakorló feladatok a *Pain_medication.sav* minta SPSS adatbázishoz készültek. Az egyes megfigyelések különböző betegek, akiken egy gyógykezelés eredményessége vizsgálható: a *status* változó 1, ha hatott a kezelés és 0, ha nem hatott a megfigyelt időszak alatt (cenzorált eset). A *time* változó az előbbi két esetre rendre a gyógyszer hatásáig eltelt idő, avagy a megfigyelési időtáv. Ezeken kívül a betegekre, illetve a kezelésre vonatkozó adatokat láthatunk, amelyek kategóriáiból almintákat képezhetünk majd. Előbbiek alapján világos, hogy az adatbázishoz kapcsolódó feladatokban a „túlélés" azt fogja jelenteni, hogy még nem következett be a vizsgált esemény, azaz nem hatott a betegre a gyógykezelés.

2. feladat:

Készítsük el a Kaplan-Meier féle becslését a túlélésfüggvénynek a fenti státusz- és időváltozókkal (almintákra bontás nélkül)!

a, Az adatok alapján hogyan becsülhető annak a valószínűsége, hogy egy betegnél épp 0,8 időegység (mostantól legyen nap) eltelte után hat a gyógyszer, feltéve hogy addig nem hatott? Mit ad meg ennek a valószínűségnek a komplementere?

b, Az adatok alapján hogyan becsülhető annak a valószínűsége, hogy egy betegnél legalább 0,9 nap eltelte után hat a gyógyszer (nem feltételes valószínűség!)? Milyen kapcsolatban van mindez a Survival table és a Survival function outputokkal?

c, Mennyi a túlélési idő mediánja és a kvartilisei? Hogyan látható ez a túlélési függvényen?

A feladat megoldása:

A Kaplan-Meier modell elérési útja az SPSS-ben:

Analyze → Survival → Kaplan-Meier

A megjelenő ablakban helyezzük a *time* változót a **Time**, a *status* változót a **Status** felirat alá, utána pedig a **Define Event** gombnál állíthatjuk be, hogy a Status változó mely értékeinél következett be a vizsgált esemény.

Define Event → Single value → írjunk be **1**-est (hiszen ez jelenti az esemény bekövetkezését, a kezelés hatásosságát). Amennyiben a Status változó nem bináris, akkor a Range of values vagy a List of values pontokban definiálhatjuk a modellezendő eseményt.

Options → kérjük a **Survival table(s), Mean and median survival, Quartiles** statisztikákat és a **Survival** ábrát! Végezzük el a futtatást!

a, Tekintsük a **Survival table** táblát, (avagy a *time* változó szerint növekvő sorrendbe tett megfigyeléseket az adatbázisban)! Részlet a táblából:

Survival Table

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	,600	Taken effect	.	.	1	199
2	,600	Taken effect	,990	,007	2	198
3	,700	Taken effect	,985	,009	3	197
4	,800	Taken effect	.	.	4	196
5	,800	Taken effect	,975	,011	5	195
6	,900	Taken effect	.	.	6	194
7	,900	Taken effect	.	.	7	193
8	,900	Taken effect	.	.	8	192
9	,900	Taken effect	,955	,015	9	191
10	1,000	Taken effect	.	.	10	190

2

197

A kérdéses valószínűséget jelölje $q_{0,80}$ (általános esetben q_c). A Kaplan-Meier féle becslés szerint:

$$q_{0,80} = P(T = 0,8 | T \geq 0,8) =$$

$$= \frac{\textit{Az éppen 0,80 idővel meggyógyulók száma}}{\textit{Azok száma, aktket legalább 0,80 ideig vizsgálatak}} = \frac{2}{197}.$$

A fenti képlet alkalmazásakor fontos látni, hogy a számlálóban az adott időponthoz (0,80) tartozó cenzorált megfigyeléseket nem kell számolni, a nevezőben viszont a cenzorált adatokat is figyelembe kell venni az összes lehetséges időponthoz.

$$p_{0,80} = 1 - q_{0,80} = P(T > 0,8 | T \geq 0,8) = \frac{195}{197}.$$

A komplementer esemény valószínűsége ($p_{0,80}$) nem más, mint annak esélye, hogy valakire nem hat a gyógyszer a 0,8 időpillanatban, feltéve hogy ezelőtt sem hatott rá, azaz a vizsgált esemény szempontjából „túléli" a 0,8-as időpontot feltéve, hogy előtte sem következett be az esemény.

b, A Kaplan-Meier féle becslés a feltétel nélküli $P(T \geq t)$ túlélési valószínűségeket a fent bemutatott p_τ valószínűségek szorzatainak segítségével állítja elő, ahol a τ helyére a mintából megfigyelt t-nél kisebb (nem cenzorált!) megszűnési időpontokat kell helyettesíteni.

$$P(T \geq t) = \prod_{\tau<t} p_\tau$$

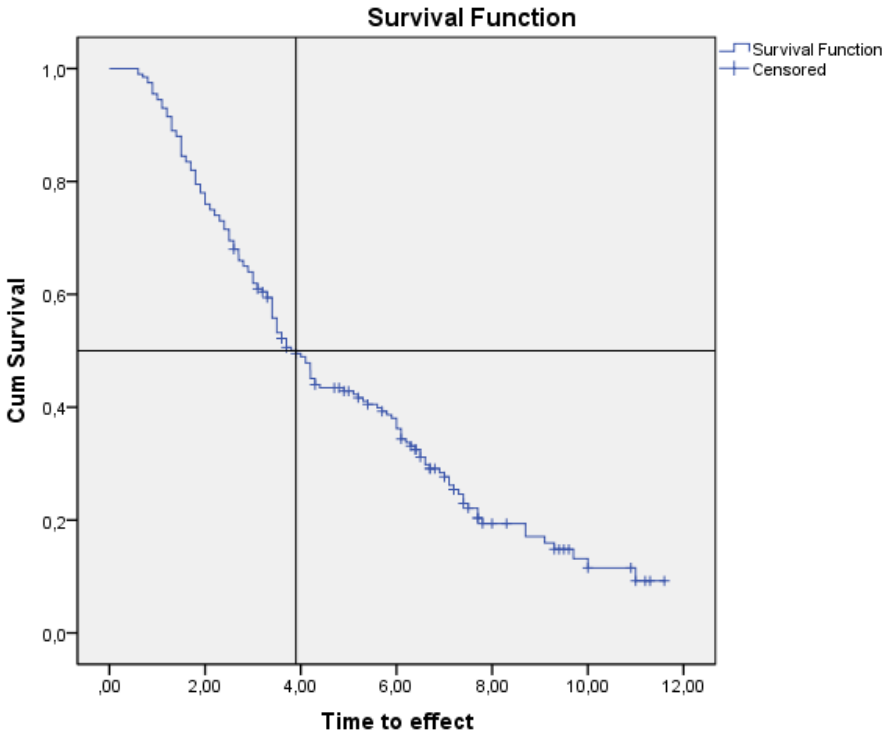
Fontos látnunk, hogy a valószínűségeloszlás becslése csak az ismert megszűnési időpontokban változik, két kilépési pont között állandó. A **Survival table** alapján 0,8 és 0,9 két szomszédos kilépési idő, ezért tetszőleges $s \in (0{,}8\,;0{,}9]$ számra igaz lesz, hogy:

$$P(T \geq 0{,}9) = P(T \geq s) =$$

$$= P(T > 0{,}8) = p_{0.60} \cdot p_{0.70} \cdot p_{0.80} = (1 - q_{0.60}) \cdot (1 - q_{0.70}) \cdot (1 - q_{0.80}) =$$

$$= \left(1 - \frac{2}{200}\right) \cdot \left(1 - \frac{1}{198}\right) \cdot \left(1 - \frac{2}{197}\right) = \frac{198}{200} \cdot \frac{197}{198} \cdot \frac{195}{197} = \frac{195}{200} = 0{,}975.$$

A kapott érték nem más, mint a **Survival Table**-ben a 0,8-as kilépési időnél szereplő becsült túlélési arány, és a hasonló módon minden t-re kiszámolható $G(t) = P(T \geq t)$ értékekből rajzolódik ki a **Survival Function** ábrán látható $G(t)$ túlélési függvény. A modell futtatásakor a **Save** menüpontban a **Survival**-t kipipálva új változóként elmenthetők az így adódó értékek.



c, A túlélési idő mediánja annak a hossza, amíg az összes megfigyelésnek a fele már kikerül a vizsgálatból. Becsült értéke most 3,9, amely könnyen látható a $G(t)$ függvény ábráján is, ott ahol a becsült túlélési valószínűség (felülről) eléri vagy átlépi a 0,5-ös értéket (lásd a $G(t)$ függvény ábráján a segédvonalakat). Hasonló logikával adódnak és szemléltethetőek a másik két kvartilis értékei is (7,3 és 2,1).
A statisztikáknál látható az átlagos túlélési idő is, ennek becslése most 5,014.

Means and Medians for Survival Time

Mean[a]				Median			
		95% Confidence Interval				95% Confidence Interval	
Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
5,014	,252	4,520	5,507	3,900	,272	3,367	4,433

a. Estimation is limited to the largest survival time if it is censored

Percentiles

25,0%		50,0%		75,0%	
Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
7,300	,371	3,900	,272	2,100	,196

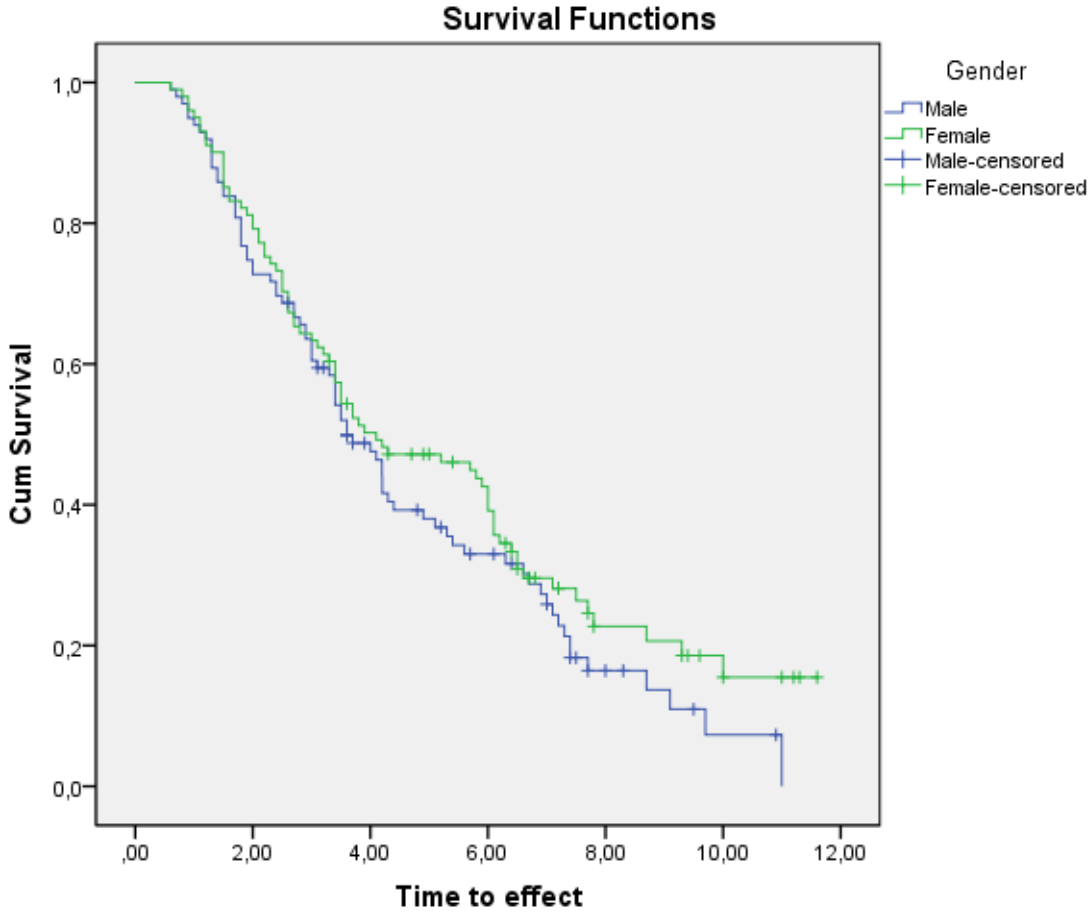
3. feladat:

Ismételjük meg a Kaplan-Meier féle becslést a *gender* változó kategóriáira képzett almintákra! Hasonlítsuk össze a férfi és női becsült túlélésfüggvényeket, az átlag és medián túlélési időket! Mit mondhatunk a túlélésfüggvények egyezéséről az SPSS által számolt próbák alapján? Érdemes ez alapján Cox-regressziót kérnünk a *gender*-rel, mint magyarázó változóval?

A feladat megoldása:

A **Kaplan-Meier** menüponton belül a **Factor** felirat alá húzzuk be a *gender* változót, valamint a **Compare Factor** gombnál kérjük a **Log rank, Breslow és Tarone-Ware** próbákat**.**

A **Survival Functions** ábrán láthatjuk a férfi és női almintákhoz tartozó becsült túlélésfüggvényeket. A megfigyelt időtáv jelentős részén közel egyező a két függvény. Két szakaszon figyelhetünk meg jelentősebbnek látszó eltérést. Például t=7 esetén a férfiakhoz kisebb függvényérték tartozik, $P(T \geq 7)$ kisebb a férfiak esetén, azaz a férfiak kisebb arányban „élik túl" ezt az időpontot, ami a konkrét példa esetén azt jelenti, hogy nagyobb arányuk esetén hatásos a kezelés eddig az időpontig, mint a nők esetén.



Az átlagos túlélési idő becsült 95%-os konfidencia intervalluma[14] a nők esetén [4,599; 6,062], a férfiak esetén pedig [4,027; 5,313]. A két intervallum átfedi egymást, így nem állíthatjuk, hogy a férfi és női átlagok jelentősen eltérnének. Hasonló logikával elemezhetjük a medián túlélési időket. A grafikonok és a becsült értékek alapján azt mondhatjuk, hogy nem teljesen egyező a férfi és női túlélés eloszlása, de nem jelentősek az eltérések.

Means and Medians for Survival Time

	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
Gender	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
Male	4,670	,328	4,027	5,313	3,600	,272	3,066	4,134
Female	5,330	,373	4,599	6,062	4,100	,922	2,294	5,906
Overall	5,014	,252	4,520	5,507	3,900	,272	3,367	4,433

a. Estimation is limited to the largest survival time if it is censored.

[14] Az átlagra adott pontbecslés a nők esetén 5,33, a 95%-os konfidencia intervallum pedig egy olyan intervallum becslése, amibe 95%-os valószínűséggel beleesik az átlag. Kiszámoljuk a férfiakra is, majd megnézzük a két intervallum metszetét, és ha átfedik egymást, akkor kicsi az esélye, hogy nagyban különbözzön a két átlag.

Az SPSS 3 tesztet számol ki a két alminta túlélésfüggvényeinek egyezésére vonatkozóan. A Mantel-Cox, Breslow és Tarone-Ware próbák nullhipotézise az, hogy a különböző kategóriák túlélésfüggvényei azonosak. 5%-os szignifikanciaszinten mindhárom teszt esetén elfogadjuk a nullhipotézist, azaz hogy egyezők a férfi és női túlélésfüggvények. Például a Mantel-Cox próba p-értéke 0,224, és mivel 0,05<0,224, ezért döntünk H_0 elfogadása mellett.

Összefoglalva azt mondhatjuk, hogy ugyan nem teljesen azonosak a férfi és női túlélésfüggvények, de a próbák alapján nem szignifikáns a köztük lévő eltérés, így nem érdemes Cox-regressziót kérnünk a *gender* magyarázó változóval (ez persze nem jelenti azt, hogy más magyarázó változókkal sem kaphatunk jó modellt).

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1,479	1	,224
Breslow (Generalized Wilcoxon)	,578	1	,447
Tarone-Ware	,825	1	,364

Test of equality of survival distributions for the different levels of Gender.

A Cox-regressziós modell alkalmazása és az eredmények értelmezése

A gyakorló feladatok a *Telco.sav* SPSS mintafájlhoz készültek. Az adatbázisban egy telekommunikációs cég ügyfeleinek (minden sor egy ügyfél) adatai láthatók. A *tenure* változó a szolgáltatónál eltöltött időt mutatja, a *churn* pedig azt hogy törölve lett-e a szerződés. Előbbi két változót rendre idő- és státuszváltozóként alkalmazva túlélési modellt kérhetünk a törlés (szerződés kezdettől mért) idejének eloszlására. A Kaplan-Meier féle becslés elkészítése a gyakorló feladatok között szerepel majd, a továbbiakban a Cox-regressziós modellt mutatjuk be az adatokon.

4. feladat:

Modellezzük a törlés idejének eloszlását Cox-regresszió segítségével! A lehetséges magyarázó változók legyenek: *age* (kor), *address* (mióta lakik a címén), *income* (jövedelem), *employ* (mióta dolgozik a munkahelyén), *gender* (nem), *wireless* (rendelkezik-e vezeték nélküli szolgáltatással), *ebill* (elektronikus számlafizetés), *internet*, *custcat* (szolgáltatás típusa – pl. Basic, Total). A Forward Wald változószelekciós eljárást használjuk! A kategorikus változókat indikátor változókként kezeljük, mindig az első kategóriához viszonyítsunk!

a, Mi a változószelekciós eljárások jelentősége? Értelmezzük, hogyan fut le a választott Forward Wald módszer, és milyen sorrendben lépteti be az eljárás a változókat! Mit mondhatunk a modell egészéről és a bevont változók szignifikanciájáról az egyes lépésekben?

b, Értelmezzük a végső modellben a regressziós együtthatókat, azok hatását a kockázati rátára és a túlélési függvényre! Mit ad meg a táblázatban az exp(B) oszlop? Hogyan számíthatók ki a bevont magyarázó változók értékei alapján adott megfigyelésre a túlélésfüggvény egyes értékei?

c, Értelmezzük a magyarázó változók átlagánál vett túlélési függvényt! Mit jelent az átlag a kategorikus változók esetén? Az ábra alapján mit mondhatunk a medián túlélési időről?

d, Rajzoltassuk le az előbbi modell szerint a *custcat* változó kategóriáira külön-külön a túlélési függvény grafikonját!

A feladat megoldása:

A Cox-regresszió elérési útja az SPSS-ben:

Analyze → Survival → Cox Regression

A **Time** és **Status** változókat állítsuk be a fentiek szerint, a **Define Event**-nél az 1-es értékét állítsuk be (ez jelenti a törlést). A **Covariates** felirat alá kell a magyarázó változókat behúzni, ezt is tegyük meg a fentiek szerint. A **Method**-nál állítsuk be a **Forward Wald** eljárást.

Kategorikus változókat 0 – 1 értékű úgynevezett dummy változók segítségével tudunk a regresszióba bevonni. Ha az eredeti változó n különböző értéket vehet fel, akkor azt $(n \quad 1)$ számú dummy változóval tudjuk helyettesíteni. Például a *custcat* változónak 4 féle értéke lehet, ezért a modellben való szerepeltetéséhez 3 új dummy változót szükséges definiálni (ezeket az SPSS elkészíti majd nekünk). A lenti táblázatban láthatunk egy lehetséges példát a kódolásra. Vegyük észre, hogy a 3 dummy változó értékei kölcsönösen egyértelműen megfeleltethetők a *custcat* változó értékének. Lényegében az történik, hogy az egyes dummy-k egy-egy kategória indikátor változói, és ha az összes dummy értéke 0, az jelzi a negyedik, kimaradó kategóriát. Ez utóbbit szokás referencia kategóriának nevezni, ehhez viszonyítjuk a többi csoportot.

	Dummy 1	Dummy 2	Dummy 3
Basic service	0	0	0
E-service	1	0	0
Plus service	0	1	0
Total service	0	0	1

Kattintsunk a **Categorical** gombra! Húzzuk be a kategorikus változókat (*gender, internet, custcat, wireless, ebill*) a jobb oldali dobozba. Alapbeállítás szerint indikátor (dummy) változók bevonásával kezeli a kategorikus változókat az SPSS. A referencia kategóriát kell átállítanunk: jelöljük ki mind az 5 változót, majd **Reference Category → First → Change**, így minden változóra az első kategória lesz a referencia érték. A futtatás után a kódolásokat a **Categorical Variable Codings** táblában láthatjuk majd:

Categorical Variable Codings[b,c,d,e,f]

		Frequency	(1)	(2)	(3)
gender[a]	0=Male	483	0		
	1=Female	517	1		
wireless[a]	0=No	704	0		
	1=Yes	296	1		
internet[a]	0=No	632	0		
	1=Yes	368	1		
ebill[a]	0=No	629	0		
	1=Yes	371	1		
custcat[a]	1=Basic service	266	0	0	0
	2=E-service	217	1	0	0
	3=Plus service	281	0	1	0
	4=Total service	236	0	0	1

A **Plots** pontban kérjük a **Survival** ábrát, majd végezzük el a futtatást!

a, A változószelekciós eljárások 3 különböző típusát különítjük el:

- **Enter módszer** – a Covariates fül alá behúzott összes változót belépteti a modellbe.
- **Forward módszerek** – egy üres, magyarázó változók nélküli modellből indul ki, majd egyesével léptet be a modellbe szignifikáns hatású változókat (és akár vesz is ki korábban bevont már nem szignifikáns változót).
- **Backward módszerek** – először az összes változót bevonja a modellbe, majd egyesével veszi ki a nem szignifikáns hatásúakat.

A Forward Wald eljárás a következőképp dolgozik. Az üres, magyarázó változót még nem tartalmazó modellből (**Block 0: Beginning Block**) indul a folyamat. A **Block 1**-ben történik a magyarázó változók lépésenkénti bevonása, a lineáris regressziónál használatos Stepwise módszerhez hasonlóan. Az ottani t-próbának esetünkben a **Wald-próba** felel meg, amelynek nullhipotézise, hogy egy adott változó magyarázó ereje nem szignifikáns (a hozzá tartozó együttható értéke 0). Így a regresszió szempontjából az az előnyös, ha „kicsi" a p-érték (a lenti táblázat **Sig.** oszlopa), ekkor érdemes szerepeltetni egy magyarázó változót a modellben. Adott egy beléptetési és kiléptetési szignifikanciaszint (alapbeállítás szerint 0,05 és 0,1, az **Options → Probability for Stepwise**-nál állíthatók). A Forward Wald folyamat lépésenkénti működése: a még nem felhasznált változók közül beveszi a modellbe a leginkább szignifikánsat (de csak azok közül, amelyeknél a t-teszt p-értéke legfeljebb a beléptetési kritériumban adott 0,05), illetve ha egy korábban már bevett változó már nem szignifikáns (p-értéke meghaladja a kilépési kritérium 0,1-es szintjét), akkor azt kiteszi a modellből (még az új változó beléptetése előtt). Az eljárás végén csak szignifikáns változók vannak a modellben, és minden kimaradó változó nem szignifikáns. A következő táblázatban az eljárás részletei láthatók, a végső modellt a *Step 8* pont mutatja. Érdekes az *age* változó: ugyan ez kerül be először a modellbe, de a 3. lépés után már nem szignifikáns, így ekkor kikerül a szereplő változók közül.

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 1	age	-,065	,006	124,361	1	,000	,937
Step 2	age	-,032	,007	22,806	1	,000	,969
	employ	-,075	,011	49,296	1	,000	,928
Step 3	age	-,002	,008	,044	1	,835	,998
	address	-,059	,010	35,184	1	,000	,942
	employ	-,080	,011	53,479	1	,000	,923
Step 4	address	-,060	,009	49,638	1	,000	,941
	employ	-,081	,010	71,408	1	,000	,922
Step 8	address	-,061	,009	47,231	1	,000	,941
	income	,002	,001	5,076	1	,024	1,002
	employ	-,079	,011	53,416	1	,000	,924
	ebill	,496	,155	10,266	1	,001	1,642
	internet	,663	,166	15,946	1	,000	1,940
	custcat			38,147	3	,000	
	custcat(1)	-1,078	,183	34,540	1	,000	,340
	custcat(2)	-,569	,197	8,339	1	,004	,566
	custcat(3)	-,799	,176	20,528	1	,000	,450

A modell egészét az úgynevezett **Omnibus** teszt minősíti, melynek nullhipotézise, hogy a bevont változók összessége nem rendelkezik szignifikáns magyarázó erővel (minden együttható 0), azaz a regressziós becslés lényegében nem jobb, mint az üres modell. Szerencsére esetünkben a p-érték 0 (lásd alább), így minden szignifikancia szinten elvethető, hogy nincs magyarázó ereje a modellnek. Ha magas az Omnibus teszt p-értéke (legalább 5%) ne használjuk a kapott modellt, keressünk más magyarázó változókat, vagy használjuk a Kaplan-Meier becslést! A végső modellhez tartozó globális teszt:

Omnibus Tests of Model Coefficients[a]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.
8	3217,781	259,714	8	,000	308,583	8	,000

a. Beginning Block Number 1. Method = Forward Stepwise (Wald)

b, A végső modell paraméterei láthatóak az alábbi táblázatban. (Ha az **Options** menüpont **Display model information** → **At last step** beállítást választjuk, akkor csak a végső modell adatait kapjuk meg.)

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 8	address	-,061	,009	47,231	1	,000	,941
	income	,002	,001	5,076	1	,024	1,002
	employ	-,079	,011	53,416	1	,000	,924
	ebill	,496	,155	10,266	1	,001	1,642
	internet	,663	,166	15,946	1	,000	1,940
	custcat			38,147	3	,000	
	custcat(1)	-1,078	,183	34,540	1	,000	,340
	custcat(2)	-,569	,197	8,339	1	,004	,566
	custcat(3)	-,799	,176	20,528	1	,000	,450

A **B** oszlopban láthatók a regressziós modell becsült együtthatói, a kockázati rátára felírt $h(t) = h_0(t) \cdot e^{b_1x_1+b_2x_2+\cdots+b_px_p}$ egyenlet b_i konstansai. $b_i > 0$ esetén, ha az x_i változó ceteris paribus növekszik, akkor növekszik a kockázati ráta (felfelé mozdul el a függvény). Ez azt jelenti, hogy növekszik az esemény bekövetkezésének feltételes valószínűsége, a túlélésé pedig csökken (a túlélési függvény lefelé mozdul el). $b_i < 0$ esetén ezzel ellentétes hatás érvényesül. Az **Exp(B)** oszlop az e^{b_i} értékeket tartalmazza. Jelentése: ha az x_i változó értéke ceteris paribus 1-gyel nő, hányszorosára változik a kockázati ráta. Az **Options** pontban konfidencia intervallumot is kérhetünk értékére.

Belátható a következő összefüggés:

$$G(t) = e^{-H(t)} = \exp(-e^{b_1x_1+b_2x_2+\cdots+b_px_p} \cdot H_0(t))$$

($H_0(t)$ az alap kumulált kockázati ráta, ez jelöli a kumulált kockázati ráta függvényt az összes magyarázó változó 0 értéke esetén)

Az **Options** pontban a **Display baseline function** négyzetet kipipálva kapjuk a **Survival table** táblázatot. Ennek **Baseline Cum Hazard** oszlopában az alap kumulált kockázati ráta, az egyes időpontokhoz tartozó $H_0(t)$ értékek láthatók. A fenti összefüggés segítségével ebből már a magyarázó változók tetszőleges értékei esetén kiszámolhatók a túlélésfüggvény pontjai. A behelyettesítésnél a folytonos változók esetén egyszerű a dolgunk, a kategorikus változóknál a **Categorical Variable Codings** kódolása alapján kell dolgoznunk (például, ha a *custcat* változó *Plus service*, akkor custcat(2)=1, és custcat(1)= custcat(3)=0). Részlet a Survival table táblából:

Survival Table

Time	Baseline Cum Hazard	At mean of covariates		
		Survival	SE	Cum Hazard
1	,025	,995	,002	,005
2	,033	,993	,002	,007
3	,072	,985	,003	,015
4	,103	,979	,004	,021
5	,144	,971	,004	,029
6	,162	,968	,005	,033
7	,186	,963	,005	,038
8	,205	,959	,005	,042
9	,235	,953	,006	,048
10	,266	,947	,006	,054

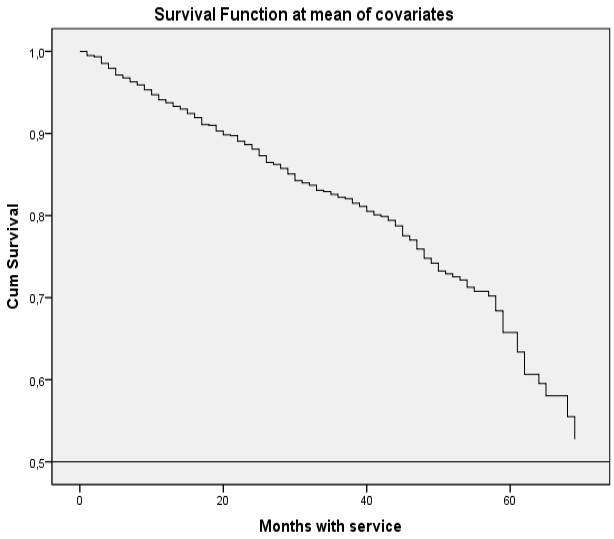
c, Az SPSS alapbeállítás szerint a magyarázó változók átlagos értékeit helyettesíti $h(t)$ regressziós egyenletébe, és az ennek megfelelő függvényeket (túlélési, kumulált kockázati ráta stb.) függvényeket rajzolja ki. Az átlagos értékek:

Covariate Means

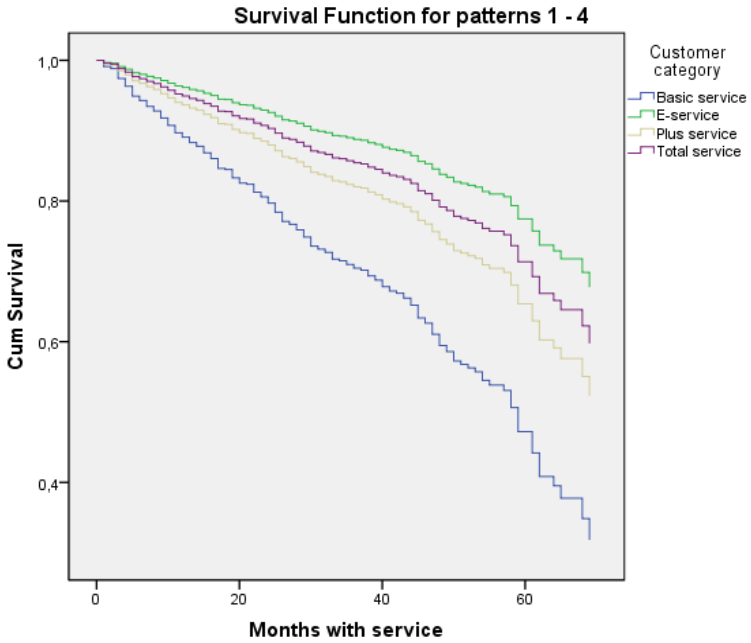
	Mean
age	41,684
address	11,551
income	77,535
employ	10,987
gender	,517
wireless	,296
ebill	,371
internet	,368
custcat(1)	,217
custcat(2)	,281
custcat(3)	,236

A dummy változók esetén a számtani átlag az 1-es érték relatív gyakorisága, így a behelyettesített becslés annyiban nem értelmes, hogy az nem vétetik fel egyik megfigyelés esetén sem. Azonban a **Plots** menüpontban az adott változót kijelölve a **Change Value → Value** mezőnél megadhatunk másmilyen választott értéket, amivel kérhetjük az ábrázolást (ne felejtsünk el a **Change-**re kattintani!).

Alább láthatjuk a túlélési függvényt. A grafikon nem metszi el a 0,5-ös szintet, így a medián túlélési idő nem becsülhető az adatok alapján.



d, Lépjünk a **Plots** menübe! Jelöljük ki a *custcat* változót, és húzzuk be a **Separate Lines for** dobozba! Újra futtatva a modellt a korábbi túlélési függvény helyett külön láthatjuk a négy kategóriához tartozó grafikonokat. A leggyorsabban a *Basic service* szolgáltatással rendelkező szerződések, leglassabban pedig az *E-service* tulajdonosok törlődnek.



5. feladat:

Ellenőrizzük a Cox regresszió mögött húzódó arányossági feltevést:

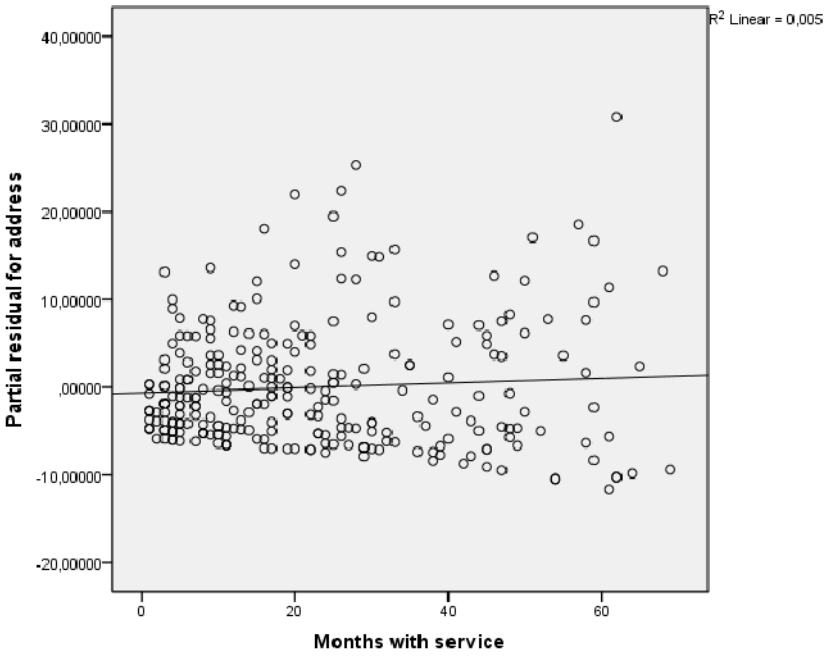
a, az *address* folytonos változóra,

b, a *custcat* kategorikus változóra!

A feladat megoldása:

Mindkét esetben a *Pénzügyi adatok statisztikai elemzése* című tankönyv Vékás Péter által írt 9. fejezetében bemutatott módszereket alkalmazzuk.

a, A **Save** menüben kérjük a **Partial residuals** pontot, majd futassuk újra a modellt. Ekkor az SPSS minden magyarázó változóhoz elmenteni az adatbázisban a hozzá tartozó úgynevezett parciális reziduálisokat (új oszlopok jelennek meg a táblában). Ezeket az új változókat az időváltozó (most: *tenure*) függvényében ábrázolhatjuk pontdiagramokon. Amennyiben a diagramon nincs trendhatás, akkor elfogadhatjuk az arányossági feltevést a vizsgált változóra. A lenti ábrán az *address* változó parciális reziduálisaira láthatjuk mindezt. Trendhatást nem mutatnak a pontok, így az arányossági feltételezés elfogadható.



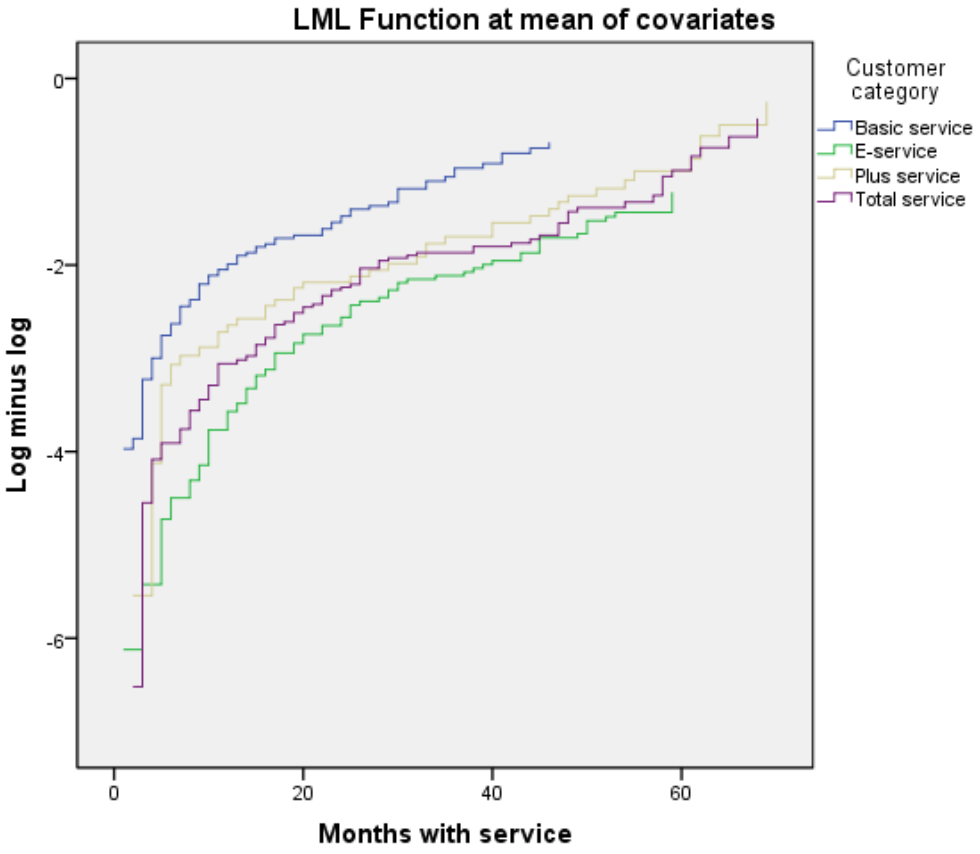
b, Vegyük ki a *custcat* változót a magyarázó változók közül, és húzzuk be a **Strata** felirat alá! Ekkor a *custcat* nem kerül be a magyarázó változók közé, a visszamaradó változókból készül a regressziós modell, úgy hogy a *custcat* kategóriái szerinti almintákra külön-külön becslést ad az eljárás az alap kockázati rátára[15] (és ezáltal az abból származtatható függvényekre is). Fontos látni a különbséget a **4. d,** feladat becslése és a mostani közt. Előbbinél a *custcat* magyarázó változó, és az egy darab regressziós egyenletbe behelyettesítve a 4 féle lehetséges értékét rajzolódik ki a 4 darab függvény. Utóbbinál pedig a *custcat* nem szerepel magyarázó változóként, hanem a 4 féle értékére külön-külön alap kockázati rátát becsül a modell. Ha ezekre elfogadható az arányossági feltevés, akkor érdemes bevenni a Cox-regresszióba magyarázó változónak a *custcat*-ot.

[15] Megjegyzendő viszont, hogy a bevont magyarázó változók és azok regressziós együtthatói ugyanazok lesznek az egyes kategóriákhoz tartozó becslés esetén.

Nem esett szó eddig az úgynevezett log minus log függvényről, ami a túlélésfüggvényből származtatható az $ln(-ln(G(t))$ képlettel. Ha az egyes kategóriákhoz tartozó log minus log görbék párhuzamosak, akkor elfogadhatjuk az arányossági feltevést.

Kérjük a **Plots → Log minus log** ábrát, ekkor megkapjuk a kérdéses görbéket. Az ábra alapján közel párhuzamosak a grafikonok, ezért elfogadható az arányossági feltevés, és a *custcat* változó modellben való szerepeltetése.

Az **a,** és **b,** feladatban láttuk, hogy az *address* és a *custcat* változók megfelelőek a modell szempontjából, de természetesen mindezt ellenőrizni kell a többi folytonos és kategorikus magyarázó változóra is a bemutatott módszerekkel.



Gyakorló feladatok

1. Adjuk meg a bevezetésben szereplő 2., 3. és 4. példák esetén a modellezendő eseményt, a túlélés és cenzorált megfigyelés jelentését, a T, s_i és t_i definícióját!

2. Adott esemény bekövetkezésére vonatkozóan adottak a következő megfigyelések:

Megfigyelés	Idő	Státusz
1	1	0
2	1	1
3	2	1
4	3	0
5	3	1
6	3	1
7	5	0
8	6	1

   Végezzük el (papíron) a túlélésfüggvény Kaplan-Meier féle becslését, majd ellenőrizzük az eredményeket az SPSS segítségével!

3. **a,** Készítsük el a Kaplan-Meier féle becslést a *Telco.sav* adatbázis *tenure* és *churn* változóira (először a teljes mintára)!

   b, Készítsünk ezután kategorikus változót az *address*-ből a Visual binning segítségével, úgy hogy az új változónak 4 lehetséges értéke legyen, és az *address* kvartilisei szerint ossza egyenlő csoportokba a megfigyeléseket! Készítsük el ezen változó kategóriái szerinti almintákra is a becslést! Szignifikáns az eltérés az egyes csoportok túlélésfüggvényei közt? A grafikonok alapján hogyan befolyásolja az *address* változó értéke a túlélést? Vonható-e párhuzam előbbi tendencia és a korábbi Cox regressziós modell (**4. b,** feladat) *address* változóra becsült együtthatója közt?

   c, Válaszoljuk meg a **b,** feladat kérdéseit az *internet* kategorikus változóból képzett alminták esetén is! (Itt nem kell Visual binning.)

4. Tekintsük a korábbi **4.** feladat Cox regressziós modelljét a *Telco.sav* adatbázisra. Végezzük el a becslést az **Enter** és a **Backward Wald** változószelekciós eljárásokkal is! Értékeljük ezeket a modelleket a korábbi szempontok szerint!

5. A korábbi **5.** feladat nyomán végezzük el a végső regressziós modellben szereplő, nem vizsgált magyarázó változókra (*income*, *employ*, *ebill*, *internet*) az arányossági feltevés teljesülését!

6. **a,** Vizsgáljuk a *Pain_medication* adatbázisra bemutatott Kaplan-Meier túlélési modellt! Ismételjük meg a becslést a *General health*, a *Treatment* és a *Dosage* kategorikus változók szerinti almintákra (külön – külön)! Tapasztalunk valahol szignifikáns eltérést az alminták túléléseinek eloszlásában?

b, Kérjünk Cox-regressziót az előbbi modell helyett, a lehetséges magyarázó változók legyenek: *age*, *gender*, *health*, *treatment*, *dosage*. A Forward Wald módszert használjuk! Mit tapasztalunk? Miért „tűnik el" a Block 1 rész az outputból? Mi történik, ha az Enter módszert használjuk? Értékeljük ezt a modellt is!

Irodalomjegyzék

Vékás Péter [2011]: *Túlélési modellek*
Megjelent: Kovács Erzsébet [2011]: *Pénzügyi adatok statisztikai elemzése,* Tanszék Kft., Budapest

Ellenőrző tesztkérdések

Jelölje be a helyes választ a következő kérdéseknél!

1. Egy gép elromlásáig eltelt időt vizsgáljuk túlélési modell segítségével. Az egyik gépet 2010.01.01-én üzemelték be és 2014.01.01-ig (a megfigyelési időszak végéig) nem romlott el. Melyik állítás igaz erre a megfigyelésre?

a) Kizárjuk az adatok közül, mert nem következett be vizsgált esemény.
b) A státuszváltozó értéke 0, az időváltozó értéke 4 év lesz.
c) A státuszváltozó értéke 1, az időváltozó értéke 4 év lesz.
d) Egyik korábbi válasz sem igaz.

2. Kaplan-Meier modellt alkalmazunk az előző feladat problémájára. Valamely kategorikus változó almintái esetén is elvégezzük a becslést, és a Mantel-Cox próba p-értékére 0,002-t kapunk.

a) A kategorikus változót érdemes lehet bevonni egy Cox-regressziós elemzésbe magyarázó változóként.
b) Az alminták túlélésfüggvényei azonosnak tekinthetők a szokásos szignifikanciaszintek mellett.
c) Mindkettő előbbi válasz igaz.
d) Egyik korábbi válasz sem igaz.

3. Egy Cox-regressziós modellben valamely x_i magyarázó változó becsült együtthatója: -0,2. Az x_i változó értékét növeljük a többi változó változatlansága mellett. Hogyan változik ennek hatására a kockázati ráta és a túlélésfüggvény?

a) Mindkettő növekszik (felfelé mozdulnak el a függvények).
b) Mindkettő csökken (lefelé mozdulnak el a függvények).
c) A kockázati ráta csökken, a túlélésfüggvény növekszik.
d) A kockázati ráta növekszik, a túlélésfüggvény csökken.

4. Egy Cox-regressziós modellben valamely x_i magyarázó változó becsült együtthatója: 0,3. Az x_i változó értékét növeljük a többi változó változatlansága mellett. Hogyan változik ennek hatására a kumulált kockázati ráta és az alap kockázati ráta?

a) Mindkettő növekszik (felfelé mozdulnak el a függvények).
b) A kumulált kockázati ráta növekszik, az alap kockázati ráta csökken.
c) A kumulált kockázati ráta nem változik, az alap kockázati ráta nő.
d) A kumulált kockázati ráta nő, az alap kockázati ráta nem változik.