

OTDK-DOLGOZAT

2015

**A LAKOSSÁGI FOGYASZTÓK MAGATARTÁSÁNAK VIZSGÁLATA
A MAGYAR VILLAMOSENERGIA-SZEKTORBAN**

**EXAMINING OF THE BEHAVIOUR OF HOUSEHOLD CONSUMERS
IN THE HUNGARIAN ELECTRICITY SECTOR**

Kézirat lezárva: 2014. március 22.

TARTALOMJEGYZÉK

1. BEVEZETÉS	1
2. A MAGYAR VILLAMOSENERGIA-SZEGMENS ÉS A SZOLGÁLTATÓ VÁLTÁS HELYZETE 3	
2.1. A LAKOSSÁGI FOGYASZTÓK HELYE MAGYAR VILLAMOSENERGIA-SZEGMENSBEN ...	3
2.2. SZOLGÁLTATÓ VÁLTÁSI HAJLANDÓSÁGGAL KAPCSOLATOS EDDIG EREDMÉNYEK RÖVIDEN	5
2.3. KÉRDÉS FELVETÉS, KUTATÁSOM CÉLJA, HIPOTÉZIS ÚJBÓLI MEGFOGALMAZÁSA.....	8
3. A LAKOSSÁGI FOGYASZTÓI MAGATARTÁS ELEMZÉSÉNEK ESZKÖZEI	10
3.1. FELHASZNÁLT STATISZTIKAI MÓDSZEREK	10
3.1.1. Alapjellemzők, függetlenség vizsgálat.....	10
3.1.2. Logisztikus regressziós modell háttere, származtatása.....	13
3.1.3. A modell illeszkedésének jóságát vizsgáló mutatók	18
3.2. AZ R PROGRAM	22
3.2.1. R objektum orientált programnyelv	23
3.2.2. Az R programnyelv gyakorlati alkalmazása a dolgozatban felállított hipotézisre	23
4. MODELL ÉPÍTÉS, TÉNYLEGES ELEMZÉS.....	31
4.1. ADATOK	31
4.1.1. Saját survey-en alapuló minta kisminta	31
4.1.2. Társadalomkutató Intézet segítségével készült minta	31
4.1.3. Magyar Energetikai és Közmű-szabályozási Hivataltól felmérésén alapuló minta	33
4.2. KISMINTÁS SURVEY	33
4.3. TÁRKI ADATBÁZISON ALAPULÓ MODELL.....	34
4.4. ÚJ MODELLEK FEJLESZTÉSE AZ OPTIMALIZÁLT ALGORITMUS SEGÍTSÉGÉVEL	42
4.5. MEKH-ES ADATBÁZISON ALAPULÓ MODELL	48
4.6. TANULSÁG, KÖVETKEZTETÉS.....	52
5. ÖSSZEGZÉS	55
IRODALOMJEGYZÉK.....	57

Megjegyzés: a Melléklet a kiegészítő dokumentumban található a feltöltési előírások miatt

MELLÉKLET	I
ALAPVETŐ R FUNKCIÓK	I
TARKI.....	VIII
Kód outputok	viii
Diagramok.....	xv
Szignifikanciakereső processzor.....	xxi
Szignifikanciakeresés eredménye	xxiv
MEKH	LXIX

<i>Kód outputok</i>	<i>lxix</i>
<i>Diagramok.....</i>	<i>lxxii</i>
<i>Feltételes valószínűségek</i>	<i>lxxiii</i>
<i>Szignifikanciakeresés eredménye</i>	<i>xcvi</i>
KÉRDŐÍVEK.....	CXIII
<i>MEKH Kérdőív.....</i>	<i>cxiii</i>
<i>TÁRKI Kérdőív.....</i>	<i>cxxvii</i>
<i>Kismintás Surveyhez tartozó kérdőív.....</i>	<i>cviii</i>

ÁBRAJEGYZÉK

1. ábra: Az egyetemes szolgáltatás területi megoszlása.....	4
2. ábra: Függetlenség vizsgálat, gyakorisági tábla.....	12
3. ábra: Kontingencia tábla (Confusion Matrix).....	20
4. ábra: 5 diszkrét klasszifikáció	21
5. ábra: Fontos Önnek a kapcsolt áruszolgáltatás?	34
6. ábra: Váltana szolgáltatót?	35
7. ábra: Átkódolás előtt az Dummy_ob413a Y.....	37
8. ábra: Előfordult-e már, hogy pénzhány miatt nem tudták befizetni az alábbi tételeket?..	39
9. ábra: Feltételes valószínűségek megjelenítése, Logit1 modell	41
10. ábra: Logit1 modell ROC görbéje és a görbe alatti terület	42
11. ábra: A feltételes valószínűségek ábrázolása hisztogramon	43
12. ábra: A változók határhozzájárulása	44
13. ábra: Feltételes valószínűségek megjelenítése, Logit2 modell	46
14. ábra: Feltételes valószínűség megjelenítése 2, Logit2 modell.....	47
15. ábra: ROC görbe és a görbe alatti terület, Logit2 modell	48
16. ábra: Eredményváltozó kimeneteleinek megoszlása.....	49
17. ábra: ROC görbe és görbe alatti terület, Logit3 modell.....	50
18. ábra: A szolgáltató váltás kérdése a férfiak körében, Logit3 modell.....	51
19. ábra: A szolgáltató váltás kérdése a nők körében, Logit3 modell	52

ABSZTRAKT

A hibrid villamosenergia-szegmensben a lakossági fogyasztók részéről nem történik számottevő szolgáltató váltás. Az egyetemes szolgáltatók területi monopóliuma a közüzemek jogutódjaként megmaradt, ami a liberalizációs célkitűzéseknek ellentmond. A szegmens látszatintézkedésektől zsúfolt. Céлом az egyetemes szolgáltatásra jogosult lakossági fogyasztók viselkedésének elemzése, hogy a problémát a fogyasztói oldal szemszögéből lássam. Hipotézisem gyakorlati tapasztalataim alapján az, hogy az alapvető demográfia változók hatással vannak a szolgáltató váltási hajlandóságra, amit a szegmensben tevékenykedő engedélyeseknek tanácsos beépíteni stratégiájukba, figyelembe véve a tömegbázis jelentőségét. Továbbá egyéb háttérváltozók szerepét is kutatom az R programnyelvet használva. Hipotézisemet bináris logisztikus regressziós modell felépítésével bizonyítom, többek között egy Tárkis reprezentatív mintán, amelynek elkészítésében én is részt vettem. A modell jóságát is alátámasztom, továbbá az egyetemes szolgáltatók közötti különbségeket is bemutatom fogyasztói elégedettséget és szolgáltató váltási hajlandóságot tekintve, mivel úgy sejttem, hogy területi ismérv szerinti eltérések vannak a fogyasztók között, amit a kismintás survey-ekben is megmutatnak. A probléma megoldását jelentheti a lakossági fogyasztóknak az egyetemes szolgáltató vállalatok versenypiaci leányvállalatának ajánlataival történő kicsalogatása az egyetemes szolgáltatásból a verseny kialakulása érdekében.

1. BEVEZETÉS

A villamosenergia-szegmensben végbement piacnyitás után 2008-tól a lakossági fogyasztók is szabadon válthatnak szolgáltatót Magyarországon a liberalizációs Európai Unió törekvések eredményeképp. A villamosenergia-szegmens egyedülállósága az egyszerre egymás mellett létező versenypiac és az egyetemes szolgáltatás kettőssége. A versenypiacon bárki vásárolhat áramot, míg az egyetemes szolgáltatásban csak a lakossági fogyasztók. A dolgozatomban a lakossági fogyasztók magatartását elemzem, vállalati fogyasztókra nem térek ki.

Az egyetemes szolgáltatásra jogosult lakossági fogyasztók körében a szolgáltató váltás elhanyagolható az ezzel kapcsolatos intézkedések ellenére is. Eddigi munkáimban az egyetemes szolgáltatás jogintézmény alakulását, értelmét, jövőbeni lehetőségeit vizsgáltam szabályozói oldalról, most célom a lakossági magatartás vizsgálata, azaz a fogyasztói oldal górcső alá vétele. Nem célom a rezsicsökkentés hatásait bemutatni, mivel ez a téma egy másik dolgozatot ölelne fel, valamint a rezsicsökkentés hatásai minden egyetemes szolgáltatót közel azonos arányban érintettek hazánkban, ezért a közel azonos arányú árváltozások miatt, és a téma mélysége miatt ettől eltekintek. Továbbá nem minden adatbázist érint a kérdés.

A szolgáltató váltási hajlandóság és általában véve a villamosenergia-lakossági fogyasztók magatartásának vizsgálatának szakirodalma legjobb tudomásom szerint nem túl tág, nem beszélve a magyar specifikusságokról. Ezért is kiemelkedően fontosnak tartom a kérdés körbejárását, elemző és kvantitatív vizsgálatát. A korábbi munkáimra, kutatásaimra és szakmai gyakorlatomra támaszkodva az a hipotézisem, hogy az alapvető demográfiai változók hatással vannak arra a kérdésre, hogy a lakossági fogyasztó szándékozik-e villamosenergia-szolgáltatót váltani. Ilyen demográfiai változók például a nem, kor, megye, lakhely, iskolázottság, anyagi helyzet, egyéb családi körülményekkel kapcsolatos jellemzők.

A dolgozatban először a kutatási kérdés háttérét mutatom be röviden, majd a hipotézisemet fogalmazom meg, ezután az elemzési eszközöket ismertetem, végül a tényleges modell építés és gyakorlati elemzés segítségével bizonyítom hipotézisem. A modellépítést R program segítségével végzem, amely egy grafikus felülettel nem rendelkező egyedi objektum orientált programnyelvet kívánó program, ami tehát programkódokon alapul. Használata világszerte elterjedt, ám hazánkban még nem jellemző. Azért választottam az R programozást a modellépítéshez, mert azon kívül, hogy gyors, viszonylag gyorsan írható benne egyedi funkció, rendkívül kompatibilis és számos új távlatot nyit meg a megszokott programokkal

ellentétben az egyedi programnyelv miatt; az R használata ritka hazánkban, így ezzel szintén szeretném bizonyítani a hasznosságát és kiemelkedő jelentőségét.

A kutatási kérdésem és a hipotézis szintén egyedinek mondható, hiszen a lakossági magatartás vizsgálata a villamosenergia-szektorban, kiemelve a szolgáltató váltási hajlandóságot nem elterjedt publikációs szakirodalmi téma hazánkban. Ezért elengedhetetlennek tartom, hogy kutatásommal felhívjam a figyelmet a téma jelentőségére, ami a lakossági fogyasztók és a villamosenergia-szolgáltatók részére is hasznosnak bizonyulhat.

2. A MAGYAR VILLAMOSENERGIA-SZEGMENS ÉS A SZOLGÁLTATÓ VÁLTÁS HELYZETE

2.1. A LAKOSSÁGI FOGYASZTÓK HELYE MAGYAR VILLAMOSENERGIA-SZEGMENSZEN

A dolgozatomban célja, hogy a közgazdasági, statisztikai és modellezési elemzésekre támaszkodva részletes képet adjak a lakossági fogyasztók viselkedéséről, ezért először az alapvető fogalmak definiálása szükséges.

A villamosenergia-szegmens Magyarországon két részre bontható, a versenypiacra és az egyetemes szolgáltatásra. A versenypiacon az árak nem reguláltak, míg az egyetemes szolgáltatásban igen. Egyetemes szolgáltatás egy speciális villamosenergia-értékesítési forma, amely átlátható, méltányos, összehasonlítható áron biztosítja az áramot az arra jogosult lakossági és bizonyos esetekben egyéb kisfogyasztóknak.¹ Az egyetemes szolgáltatás a közszolgáltatásokból azért fejlődött ki, mivel bizonyos szolgáltatásokat minden fogyasztó számára biztosítani kell, mintegy alapvető jogként, ezeket nem szabad csupán a piacra bízni. A hálózatos iparágakban ezt a szerepet többnyire az egyetemes szolgáltatás jogintézménye vette át. A lakossági fogyasztók nagy része, azon része, amely nem lépett ki a kevésbé regulált versenypiacra, az egyetemes szolgáltatásban helyezkedik el. Az egyetemes szolgáltatás részletes bemutatása egy az egyetemes szolgáltatás helyzetének vizsgálatáról szóló (2012) tanulmányomban olvasható. Így az egyetemes szolgáltatás bemutatására csak annyira térek ki, hogy a kutatási kérdés felvetésének háttere az olvasó számára világossá váljon. Így a törvényi háttér, elemzési pontok, következtetések a szerző egy korábbi (2012) munkájában láthatóak, amely az egyetemes szolgáltatás helyzetét vizsgálja. Az anonimitási kritérium értelmében a saját magamra való hivatkozást (Szerző, 2012) módon teszem meg, a helyes hivatkozási forma kielégítése érdekében, ami az előbb megemlített kutatásra utal.

Az egyetemes szolgáltatók keretszabályozását a 2003/54/EK biztosítja. A magyar szabályozásban a 2007. évi LXXXVI. villamos energia törvényben (VET) jelenik meg az új fogalom, ezzel átvéve az egykori közüzemek szerepét. A VET szerint „az egyetemes szolgáltatás: a villamosenergia-kereskedelem körébe tartozó sajátos villamosenergia-értékesítési mód, amely Magyarország területén bárhol, meghatározott minőségben a jogosult felhasználó számára méltányos, összehasonlítható, átlátható ár ellenében igénybe vehető.”

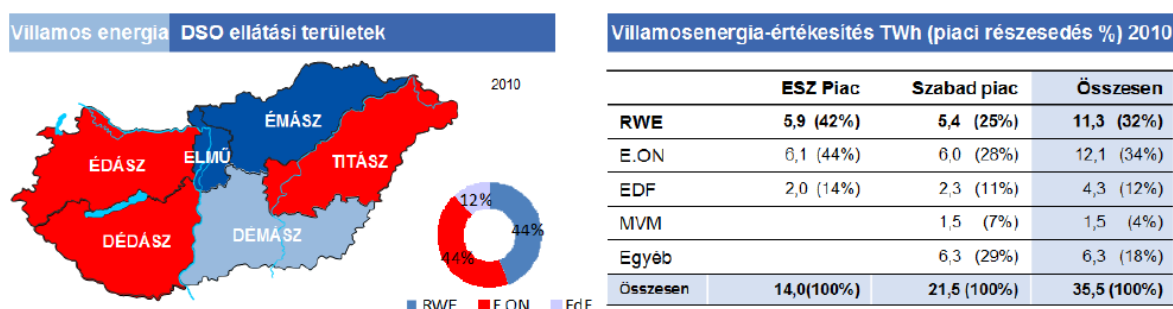
¹ Bővebben 2007. évi LXXXVI. törvény a villamos energiáról 3. 48-52. paragrafus 2001. évi CX. törvény a villamos energiáról, egységes szerkezetben a végrehajtásról szóló 180/2002. (VIII.23.) Korm. Rendelettel

2009/72/EK irányelv a villamos energia belső piaci szabályozásáról és a 2003/54/EK hatályon kívül helyezéséről, 3. cikk 273/2007 rendelet a villamos energiáról szóló 2007. évi LXXXVI. tv. egyes rendelkezéseinek végrehajtásáról

(2007. évi LXXXVI. tv, 3. paragrafus, 7. pont) Az 48.-52. paragrafus határozza meg az egyetemes szolgáltatás egyéb törvényi kereteit: kiemeli, hogy az egyetemes szolgáltató köteles villamos energiát értékesíteni és szerződést kötni az egyetemes szolgáltatásra jogosult fogyasztókkal, tehát az erre nem jogosult fogyasztókkal nem köthet szerződést. (Szerző, 2012)

Az egyetemes szolgáltatásra jogosult fogyasztók közé tartoznak a lakossági fogyasztók, valamint a 3*63 A-nál nem nagyobb csatlakozási teljesítményű fogyasztók. Ezenfelül meghatározott költségvetési szervek, közfeladatot ellátó intézmények (például egyházi jogi személy, alapítványi fenntartású intézmény) is jogosultak lehetnek az egyetemes szolgáltatásra. (2007. évi LXXXVI. tv) Tehát a lakossági fogyasztók az egyetemes szolgáltatásban helyezkednek el, számottevő szolgáltató váltás sem az egyetemes szolgáltatókon belül, sem a két fő komponens között nem történt 2008 óta.

Köztudott, hogy az áramszolgáltatók esetében nagyobb a háztartások száma, mint amit a Központi Statisztikai Hivatal a népszámlálás esetén kimutat, mivel olyan egységeket is regisztrálnak, ami nem számít statisztikailag annak, például közös helyiség. (Sugár, 2007) 2001-ben 5,1 millió háztartás volt regisztrálva az egyetemes szolgáltatásban annak ellenére, hogy a tényleges háztartások száma 4 millió alatt maradt. 2010 januárja óta egy egyetemes szolgáltató betörhet egy másik egyetemes szolgáltató területére, ahol eddig területi monopóliummal bírtak az egyetemes szolgáltatók, ahogy az 1. ábra mutatja.



1. ábra: Az egyetemes szolgáltatás területi megoszlása

Forrás: Jászay, 2010

A területi elhelyezkedés változatlan maradt, nem törték be egymás területére az egyetemes szolgáltatók, azaz úgy vélem, hogy a rendelkezés, amely lehetővé tette a behatolást egymás területére, csupán látszatintézkedés. A tömegbázis birtoklása előnyt jelenthet, viszont a nyomott árak és a szigorúan szabályozott ügyfélszolgálati és egyéb kötelezettségek miatt a

betörés nem valósult meg szolgáltatói oldalról.² Viszont a fogyasztói oldal elemzésével a szakirodalom nem foglalkozik mélyrehatóan, ezért a lakossági fogyasztók viselkedésének elemzése elengedhetetlen. A versenypiaci szolgáltatók részéről sem volt törekvés a lakossági fogyasztókért a Magyar Telekom komplex szolgáltatás csomagjának megjelenéséig ami gázt, áramot és telekommunikációs szolgáltatásokat is nyújtott. 2012-ig megközelítőleg 150 000 egyetemes szolgáltatásra jogosult lakossági fogyasztót emeltek ki az egyetemes szolgáltatásból, (Szerző, 2012) ami annak volt köszönhető, hogy az energiadíjak a szabályozott maximalizált ár alá mentek, ami csak a keresztfinanszírozás segítségével volt lehetséges. A ragadozó árazást előidéző keresztfinanszírozás tiltott, ám ezt Magyarországon egyik felügyelet sem vizsgálja, mivel nincs egy átfogó felügyeleti rendszer, valamint a szabályozó hatóságok közötti kommunikáció sem kielégítő, ami szintén befolyással lehet közvetve a lakossági fogyasztókra.

Eddigi munkáimban az egyetemes szolgáltatás jogintézményét vizsgáltam szabályozói oldalról megközelítve, ám ebben a dolgozatban a fogyasztói oldalról történő megközelítés, a lakossági fogyasztók magatartásának, viselkedésének megértése különböző statisztikai eszközökkel a fő célom.

2.2. SZOLGÁLTATÓ VÁLTÁSI HAJLANDÓSÁGGAL KAPCSOLATOS EDDIGI EREDMÉNYEK

RÖVIDEN

Ahogy már említettem a szakirodalom nem bővelkedik a lakossági fogyasztók magatartásának ilyen szintű specifikus vizsgálatával a nemzetközi villamosenergia-szegmensben sem, ezért is tekintem fontosnak a téma megismertetését, kutatását, ami a szolgáltató multinacionális vállalatok stratégiájába is beépíthető elemeket is tartalmazhat. Az alábbi fejezetben röviden térek ki a főbb megállapításokra, amelyet a gyakorlati részben is fel fogok használni.

A háztartási fogyasztók magatartását, szolgáltató váltási jellemzőit először Klemperer vizsgálta mélyebbre hatóan az 1980-as évek végén, 1990-es évek elején. Motivációját az egyre inkább elterjedt liberalizációs törekvések adták. Tanulmányaiban a lakossági fogyasztó döntése szolgáltató váltásról költség-haszon elemzésen alapul, tehát azzal az alapfeltételezéssel él, hogy a fogyasztó minden költsége pénzben kifejezhető legyen, hiszen

² Bővebben 4./2011. (I.31.) NFM rendelet a villamos energia egyetemes szolgáltatás árképzéséről 273/2007 rendelet a villamos energiáról szóló 2007. évi LXXXVI. tv. egyes rendelkezéseinek végrehajtásáról

így válik összehasonlíthatóvá. Munkája Weizsäcker's (1984) tanulmányán³ alapszik, ám ellentétes következtetésekre jutott, mivel Klemperer Weizsäckerrel ellentétben nem határozta meg, hogy minden időszakban egyenlő legyen az ár. Egy kétidőszakos játékban a vállalat sokkal jobban küzd a második időszakbeli profitért, mintha csak egyszerűen profitmaximalizálna, mivel a második időszaki profit az első időszakitól függ. Ezen fokozott elvárások miatt a lakossági fogyasztók azt feltételezhetik, hogy a második időszakban megnövekedett piaci részesedés következtében, amit az első időszak agresszív viselkedése miatt értek el, az ár növekedni fog, így a kereslet rugalmassága csökkenni fog, ami visszahat az első időszakra is, vagyis a nettó hatás kérdéssé válik. (Klemperer, 1987a) Ezen modell és a többperiódusú kiterjesztett változatnak a matematikai összefoglalására a dolgozat szűk témája miatt nem térek ki, csak a legfőbb megállapítást emelem ki.⁴

A magyar szegmenst illetően a 2008-as teljes piacnyitás előtt a fokozatos piacnyitás következtében már 2000-től találhatunk a szakirodalomban néhány elemzést, amely a jövőbeli helyzet elemzését, lehetséges kimeneteleit tűzte ki célul. Ezek az elemzések többnyire azt mondták ki, hogy a szabad szolgáltató váltással a lakossági fogyasztók tömeges versenypiacra történő kilépése fog bekövetkezni. Egy nemzetközi szinten is kiemelkedő elemzés Tóth Pálma és Szolnoki András munkája (2007), amiben a lakossági fogyasztók szolgáltató váltását és a villamosenergia-árak alakulását vizsgálták ex ante jelleggel, azaz 2007-ben modellt készítették a 2008 január 1-jétől fennálló új liberalizált villamosenergia-szegmensre. Többek között a szolgáltató váltások nagy számát jósolták,⁵ valamint megfelelően jelezték előre az árak alakulását is, amit egy korábbi dolgozatomban is megfogalmaztam. (Szerző, 2012) (Szolnoki-Tóth, 2007) A Budapesti Elektromos Művek vezetőivel készített interjúim és a lekérdezett adatok is a szolgáltató váltás alacsony számáról tanúskodnak.

Érdemes megemlíteni a témában Hetesi Erzsébet és Rekettye Gábor kutatásait A lakossági ügyfelek lojalitásának mérése és elemzése egy magyar energiaszolgáltatónál címmel. Ez a tanulmány 2001-ben készült, vagyis a piacnyitás előtt, az egyetemes szolgáltatók jogelődjeinek, a közüzemek korában. A magyar villamosenergia-szegmens átalakulása előtt előrevetítették, hogy a közüzemi szolgáltatások versenyhelyzet következtében a szervezeti struktúra átalakulása és a vállalati magatartás módosulása fog bekövetkezni, amibe beletartozik az új piacok szerzése, a fogyasztók elcsábítása a

³ Weizsäcker (1984) The Cost of Substitution, *Econometrica*, Vol 52.

⁴ Ám részletesen Klemperer (1987a) és (1987b) munkájában látható róla összefoglaló.

⁵ A tanulmány részletesebb kibontása, értékelése, összevetése a valós helyzettel a szerző (2012) és Szolnoki Pálma-Tóth András (2007) című munkájában található.

konkurenciától, a marketing részlegek fejlődése és legfőképp a fogyasztók megtartása, a lojalitás kialakulása. A kutatás kimutatta, hogy a lakossági fogyasztók kevésbé érzékenyek, ahogy erre Sugár András (2007) is rávilágított. A Hetesi-Rekettye (2001) kutatás további érdekessége, hogy a fogyasztók egy lehetséges piacnyitás esetén is hűségesek maradnának jelenlegi (akkori) szolgáltatójukhoz az évek alatt felhalmozódott esetleges negatív élmények ellenére is. Ez érdekes eredmény, viszont ennek pontos mélyebb okaira a kutatás nem ad választ. Én a gyakorlati fejezetekben megkíséreltem a megbúvó háttérváltozók felkutatását, amit választ adhat a jelenség okaira. Végül arra a következtetésre jutottak, hogy az idősebbek, alacsony iskolai végzettségűek és a rossz anyagi körülmények között élők szolgáltató váltását nem befolyásolja a szélesebb választék, egyszerűbb ügyintézés vagy egyéb könnyítő körülmények. A gyakorlati és kutatási tapasztalataim is adtak motivációt ahhoz, hogy a téma mélyebb kibontását, elemzését kíséreljem meg a dolgozatomban. Újra kiemelem, hogy kutatási célom bemutatni és megvizsgálni, hogy a demográfiai változók milyen hatással vannak a fogyasztói viselkedésre, amiben a Hetesi Erzsébet által használt változók is szerepelnek.

A lakossági magatartás kutatása kapcsán mindenképp ki kell emelni a Sugár András és Kaderják Péter által vezetett Kisfogyasztók a liberalizált villamosenergia-piacon (2007) (Regionális Energiagazdasági Kutatóközpont) című tanulmányt, ami többek között a fogyasztói tudatosságokat, a fizetési problémákat, a keresleti ár rugalmasságot és a szolgáltató váltás modellezését is magában foglalja. Nem célom összefoglalni ezen nagylélegzetű tanulmány eredményeit, ám a végkövetkeztetések közül néhányat az alábbi fejezetekben fel fogok használni.

A szakirodalmi háttér kutatásáról és felhasználásról tanúbizonyságot téve köteles vagyok megemlíteni az egykori Magyar Energia Hivatal (ma Magyar Energetikai és Közmű-szabályozási Hivatal) által kötelezően készített fogyasztói elégedettségrel kapcsolatos felmérések szerepét. Ennek háttere, hogy a villamos energia termelésről, szállításról és szolgáltatásról szóló 1994. évi XLVIII. törvény 5. paragrafusában foglaltak szerint a villamosenergia-szolgáltatók tevékenységének hatósági felügyelete keretében kötelesek ellenőrizni a szolgáltatás színvonalát a fogyasztói érdekek védelmében. Ez a fogyasztói elégedettség rendszeres felméréséhez teremtette meg az alapot, amely jelenleg napjainkban is tart. A kutatások évente zajlanak, amelynek információ bázisát reprezentatív, véletlen jellegű kombinált mintavétel alapozza meg. A vizsgálat fő kutatási témakörei a szolgáltatás minősége, az üzemszerű kapcsolattartás, a fogyasztókkal való kommunikáció, az árak

területe és az áramszolgáltatás megítélése egyéb közszolgáltatásokhoz (víz, gáz stb.) voltak. Az alapkoncepción azon a nemzetközi elven nyugszik, miszerint a szolgáltatások egyes összetevőivel való elégedettség a legcélszerűbben oly módon értékelhető, ha ezen komponenseket a fogyasztók nekik tulajdonított fontosságához viszonyítják. A gap analízis (rés vizsgálat) szerint a fontosság/elégedettség optimális arány akkor áll fent, ha a fontossági és az elégedettségi indexek azonosak, mivel ha a fontosság meghaladja az elégedettséget, akkor alulteljesítésről, még ellenkező esetben túlteljesítésről beszélhetünk. A marketing szempontjából az enyhe túlteljesítés jelent optimális megoldást. Az 1990-es évek végén alulteljesítés jellemezte az áramszolgáltatás minőségét, még az üzemszerű kapcsolatok és a fogyasztókkal való kommunikáció túlteljesített. Az eredmények a strukturális különbségekre is engednek következtetni. Több évben is felhívták a figyelmet a szerzők arra, hogy az elégedettségi mutatók a lakossági fogyasztók iskolai végzettsége és lakóhelynek jellege szerint eltérnek, ami még inkább arra adott okot, hogy én is vizsgálódjak napjainkban a témában. Az iskolai végzettség növekedésével növekednek az áramszolgáltatással szembeni elvárások, és csökken az ugyanazon teljesítménnyel való elégedettség. A lakóhely szerint is eltéréseket tapasztalhatunk 1996-1998 között, ugyanis a kisebb városokban (azaz település típusonként eltérően) jobbnak értékelik a szolgáltatást, mint a nagyobb városokban. (Rekettye, Orosdy, Tersztyánszky, 1996-2004)

Ezen a fogyasztói elégedettség vizsgálatban is megjelenő demográfiai különbségek is okot adtak a hipotézisem megfogalmazására, aminek bizonyítása a háttérváltozók kutatásával és modellépítéssel a gyakorlati fejezetben látható.⁶

2.3. KÉRDÉS FELVETÉS, KUTATÁSOM CÉLJA, HIPOTÉZIS ÚJBÓLI MEGFOGALMAZÁSA

Miután az olvasó alapvetően megismerkedett a magyar hibrid villamosenergia-szektorral nyilvánvalóvá vált a téma fontossága. Tehát a lakossági fogyasztók részéről nincs jelentős szolgáltató váltás, amelynek az okai nem pontosan tisztázottak különböző elemzési módszerekkel sem. A fogyasztóknak 27%-a nem tudta, hogy képes szolgáltatót váltani a villamosenergia-szegmensben, az egyik adatbázisom alapján, amely reprezentatívnak tekinthető. (Pap, 2013) Tehát a háztartási fogyasztók döntő többsége tisztában van a szolgáltató váltás lehetőségével, ám még sem él vele. Vajon akkor mi lehet a befolyásoló erő? Munkámban ezt kutatom.

⁶ A fogyasztói magatartás módszertani vizsgálatához szintén segítséget adott Pintér József és Rekettye Gábor (2008) A villamosenergia-szolgáltatással való elégedettség mérésekor alkalmazható kvantitatív módszerek és modellek című tanulmánya.

Úgy vélem, hogy a fogyasztók magatartását az alapvető demográfiai változók is befolyásolják, amire már Hetesi Erzsébet is felhívta a figyelmet, ám a piacnyitás előtti időszakra érve. Jelen dolgozatban célom e hipotézisem bizonyítása 2013-as és 2009-es adatokat felhasználva, valamint egyéb magatartásbeli jelenségeket is szándékomban áll elemezni, ami a szakirodalom szűkössége és az adatok nehéz elérhetősége miatt, valamint amiatt, hogy R programnyelvet használok a modellezéshez Magyarországon egyedülállónak számíthat.

3. A LAKOSSÁGI FOGYASZTÓI MAGATARTÁS ELEMZÉSÉNEK ESZKÖZEI

3.1. FELHASZNÁLT STATISZTIKAI MÓDSZEREK

A statisztikai módszerek lényegre törő megismertetése a dolgozat gyakorlati szempontjának alátámasztása céljából elengedhetetlen, ezért az alábbiakban erre kerül sor.

3.1.1. Alapjellemzők, függetlenség vizsgálat

Mielőtt kiválasztjuk a megfelelő adatelemzési módszert a többváltozós adatelemzés módszerei közül érdemes a változóinkat egyesével megvizsgálni. Kétféle megközelítést tartok szem előtt: egyrészt, hogy milyen válaszokat adtak a lekérdezettek, amelyeket egydimenziós elemzésekkel vizsgálunk, másrészt bizonyos változókat milyen háttérváltozók befolyásolhatnak, aminél többdimenziós elemzéseket és kapcsolat vizsgálatot végzünk. A módszer mélyebb megismertetése szükséges ahhoz, hogy megfelelő biztonsággal használjuk a módszert. A leíró statisztikáknál az átlagon, szóráson, variáncián, minimumon, maximumon, terjedelmen, mediánon és móduszon túl vizsgálhatjuk (Kovács, 2006 és Hunyadi, 2010 alapján):

- az átlag sztenderd hibáját $\frac{\sigma}{\sqrt{n}}$, az átlagok különbségének értékét méri, ami a minta nagyságának növelésével csökken;
- a ferdeség mértékét: $\gamma_1 = \frac{\sum(x_i - m)^3}{\sigma^3}$, azaz a harmadik centrális momentum és a szórás köbének a hányadosa, ahol a negatív érték jobbra elnyúló eloszlást mutat. A ferdeség és az átlag sztenderd hibájának hányadosa adja a t-próba összefüggését;
- a csúcsosság mérőszámát: $\gamma_2 = \frac{\sum(x_i - m)^4}{\sigma^4}$, ahol a negatív érték a normális eloszlásnál laposabb eloszlást jelent. A minimális értéke -2 lehet.

Az outlierok vagyis a kiugró értékek veszélyt jelenthetnek az elemzésünkre nézve, amikkel számolnunk kell. Ilyenkor alkalmazhatunk interkvartilis terjedelmet mutató ábrákat, nyesett átlagot, valamint levél ábrát is.

A normalitási vizsgálatot Kolmogorov-Szmirnov próba elvégzésével tesztelhetjük, amely hipotézis vizsgálaton alapszik. A hipotézis vizsgálat esetén a null hipotézisban (H_0) azt tesszük fel, hogy a vizsgált eloszlás normális eloszlású, azaz $F(y) = F(y_0)$, ahol az $F(y)$ folytonos eloszlás. még a H_1 hipotézisban ennek az ellentettjét tesszük fel. A teszt próbafüggvénye $D_n = \max|F_n(y) - F_0(y)|$, ahol az $F_n(y)$ maga az empirikus eloszlás függvény. A D_n miatt a próba felső kritikus tartománnyal rendelkezik. A próba

végrehajtásához a kritikus értékeket a KS táblázatból kell kikeresni. A grafikus normalitás vizsgálatához a Q-Q (quantile-quantile) teszt alkalmazható. (Hunyadi, 2002) Az alapadatokat sztenderdizálása után az empirikus eloszlást függvényt kell elkészítenünk, amit a normális eloszlás függvény inverze szerint transzformálunk, maradék változónál: $u = \Phi^{-1} \left[\Phi \left(\frac{x - \bar{x}}{s} \right) \right]$. Ha az eredeti eloszlás normális volt, akkor a kiinduló adatok oda-vissza transzformált értékei megegyeznek, vagyis az így nyert koordináták egy egyenesen helyezkednek el. Ha a változó közel normális eloszlást követ, akkor 45°-os egyenes körül szóródnak. (Hunyadi, 2012ab)

A logisztikus regressziós modell felépítéséhez szükséges a változók közötti kapcsolat feltérképezése, hogy létezik-e kapcsolat, illetve milyen erősségű, vagy irányú a fennálló kapcsolat. A változók közötti kapcsolatot vizsgáló módszerek alkalmazása a mérési skáláktól is függ. Az alapvető mérési skálák a következők (Hunyadi, 2010):

- Nominális mérési skála: megállapítható, hogy az ismértértékek egyezők vagy nem, erre példa a lakhely típusa változóm (főváros, város, község);
- Ordinális mérési skála: az ismértértékek közötti különbség plusz információt hordoz (sorrendiség), például érdemjegyek;
- Intervallum mérési skála: az ismértértékek különbözősége hordoz többlet információt, például hőmérséklet, a nulla pont önkényes;
- Arány mérési skála: az ismértértékek arányait értelmezzük, a nulla pont nem önkényes, például jövedelemmel kapcsolatos kimutatások.

A csak nominális vagy nominális és ordinális (vegyes) skálán mért változókra függetlenségi vizsgálatot végezhetünk, még az ordinális skálájú változókra rangkorrelációt alkalmazunk, az intervallum vagy arány skálán mért változókra pedig korrelációt számolunk. Keresztábra elemzéskor két változó közötti függetlenség fennállásának hipotézisét mérjük, azaz, hogy van-e valamilyen jellegű kapcsolat a két változó között. Ez a vizsgálat nem mondja meg, hogy milyen erős és milyen irányú kapcsolat van a két változó között. Ha a H_0 hipotézist elvetjük, mivel a p érték nulla, akkor a két változó független egymástól. A függetlenség vizsgálat elméleti alapját két nominális vagy ordinális skálán mért változó esetén az együttes előfordulásuk gyakoriságainak keresztábrába való rendezése adja.

	1	...	s	Összeg
1	N_{11}	...	N_{1s}	R_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	N_{r1}	...	N_{rs}	R_r
Összeg	C_1	...	C_s	n

2. ábra: Függetlenség vizsgálat, gyakorisági tábla

Forrás: Hunyadi (2010) alapján saját szerkesztés

A két ismerv szerinti osztályok az r és az s .

$$H_0: P(A_i B_j) = P(A_i)P(B_j) \quad \forall i \in \{1 \dots r\} \text{ és } \forall j \in \{1 \dots s\},$$

tehát a null hipotézis kimondja, hogy a két ismerv független egymástól. A próbafüggvény értéke általánosan felírva:

$$T = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{N_{ij}^2}{R_i C_j} - 1 \right) \sim_{H_0} \chi^2(r-1)(s-1),$$

ahol a v szabadságfok $(r-1)(s-1)$ és a kritikus érték a c_α , amely a χ^2 táblázatból olvasható le, vagyis egyoldali a próba. Ha a $T > c_\alpha$, a null hipotézist elvetjük, vagyis nem független a két vizsgált ismerv egymástól. Ha a függetlenség hipotézisét elvetjük, akkor tanácsos mérni az asszociáció szorosságát. (Kovács, 2006)

Vagyis eddig azt vizsgálta a módszertan, hogy van a kapcsolat a két ismerv között, az asszociációs mérőszámok, pedig azt mutatják meg, hogy milyen irányú a kapcsolat. Az ilyen mérőszámoknak két fajtája van: a χ^2 -en alapuló és a nem χ^2 -en alapuló mutatók. A nominális változókra számolható Phi, Cramer-V mutató és kontingencia együttható, Goodman-Krusal tau mutató, valamint Guttman féle lambda mutató, amelyek χ^2 alapú mérőszámok családjába tartoznak. A Cramer V mutató tehát nominális változók közötti asszociációt méri, eredmény képpen 0 és 1 közötti számot kapunk.

$$C = \sqrt{\frac{\chi^2}{N \min\{(r-1)(s-1)\}}}$$

ahol χ^2 a Pearson féle χ^2 , az N a megfigyelések száma. Vagyis ha közel 1 a mutató értéke, akkor erős kapcsolat van a két változó között. A Phi mutató is 0 és 1 között veheti fel értékeit

hasonló értelmezéssel. A $\phi = \sqrt{\frac{\chi^2}{n}}$, ami ritkán lehet nagyobb, mint egy, mivel a χ^2 várható értéke a szabadságfok. A lambda nem χ^2 alapú mérőszám. A GK tau számolására három változat létezik. (Barna, 2010)

A változóim között szerepelnek ordinális változók is, ezért kiemelem az ordinális változókra alkalmas mérőszámok közül a Goodman-Krusal féle Gammát:

$$\gamma = \frac{(P-Q)}{(P+Q)}, \text{ ahol } P = \sum_{i=1}^r \sum_{j=1}^c f_{ij} S_{ij} \text{ és } Q = \sum_{i=1}^r \sum_{j=1}^c f_{ij} D_{ij},$$

ahol az S az egyezően rendezett megfigyelések száma, még a D az eltérően rendezett párok száma. A P a konkordáns, azaz az egyirányú párokat jelenti, ahol az egyik rekord mindegyik változójához tartozó skálán magasabban rangsorol, mint a másik megfigyelt rekord, a Q pedig a fordított pár. Vagyis a γ megmutatja, hogy a konkordáns párok valószínűsége mennyivel nagyobb a diszkordáns párok valószínűségénél. Látható, hogy ez nem χ^2 – en alapuló mérőszám. Az értéke -1 és 1 között mozoghat, ahol a nulla jelenti a függetlenséget, ez azonban csak 2*2-es tábla esetén biztos függetlenséget. Az egyezéseket nem veszi figyelembe, vagyis kihagyja azokat az eseteket ahol egyenlőség van. A Sommer féle D mutató $D(A|B)$ azt vizsgálja figyelembe véve az egyezéseket, hogy a B1 és B2 különbözősége esetén az A1 és A2 viszonya mennyiben monoton, vagyis így az A változót tekinthetjük függő változónak. Tehát

$$D_{(A|B)} = \frac{P-Q}{D_r}, \text{ ahol } D_r = n^2 - \sum_i m_i^2, \text{ ha az oszlopban van a függő változó, még}$$

$$D_{(B|A)} = \frac{P-Q}{D_c}, \text{ ahol } D_c = n^2 - \sum_j n_j^2, \text{ ha a sorban van a függőváltozó.}$$

Mivel ez aszimmetrikus mutató, a szimmetrikus mutatót az aszimmetrikus mutató középértékeként kapjuk meg. A mutató -1 és +1 között mozog a két változó monotonitását mutatva meg. (Kovács, 2006)

3.1.2. Logisztikus regressziós modell háttér, származtatása

Az alapstatisztikák és a függetlenség vizsgálat után a változók közötti kapcsolatot megsejtve építettem fel a logisztikus regressziós modellt. A változók jellemzői miatt, amit a gyakorlati fejezetben részletesen megindokolok, nem alkalmazok multinomiális logisztikus regressziót (ahol az eredmény változónak több mint két kimenete van), hanem bináris logisztikus regresszió építését tartom célszerűnek, mivel a probléma a diszkrét döntési modellek egy speciális esete, amikor a magyarázni kívánt változó csak két értéket vehet fel.

Tehát a módszer két diszjunkt kategória bekövetkezési esélyeinek arányát, az *odds*-ot veszi alapul. A módszer gyökerei a 19. századig, P. Verhulst matematikusig nyúlnak vissza. A történelmi előzményekről jó összefoglalót ad J.S. Cramer (2002) műve. A probit modell feltételezi, hogy a p standard normális eloszlás követ, a logit modellel ellentétben (logisztikus eloszlás). Mivel a normális eloszlás eloszlásfüggvénynek nincs zárt alakja, ezért a logisztikus módszer alkalmazása egyszerűbb. A probit modell elvétve kerül alkalmazásra a gyakorlatban, származtatása összetettebb. Összességében a két módszer hasonló eredményt ad. (Tétényi, 2013)

Az eredményváltozónak ilyen esetben tehát két kimenete van:

$$Y = \begin{cases} 1 & | \quad p \\ 0 & | \quad 1 - p \end{cases}$$

Ennek a bekövetkezése nem becsülhető lineáris regresszióval, mivel nem normális, hanem binomiális eloszlású, ezért más módszerre van szükség. A logisztikus regresszió a klasszifikációs eljárások közé tartozik, mivel a rekordok egymást kölcsönösen kizáró csoportokba sorolhatóak. A bináris modellek egyik speciális esete a logisztikus regresszió. Ezzel a regresszióval végzett diszkriminálást akkor érdemes végezni, hogyha nem normális a változók együttes eloszlása. A bináris logisztikus regresszió az általánosított lineáris modellek közé (GLM), a logisztikus modelles család körébe tartozik, ám az OLS (ordinary least square) modellekkel ellentétben egy speciális loglineáris modell, ami esély hányadossal (*odds*) és feltételes valószínűségekkel (p) operál. A multikollinearitás és a variancia infláló faktor logisztikus regressziónál a reziduumok eltérésén alapuló számítás miatt nem alkalmazható. A magyarázó változók a modellben lehetnek nominális, ordinális vagy intervallum és arányskálán mért változók. (Gray et al., 2001)

A nominális magyarázó változók lehetséges kimenetelei közül egyet rögzítünk, és ezekhez viszonyítva becsüljük a Y -ra gyakorolt hatást. Így kapjuk a kovariánsokat, ami az értelmezésnél fontos. Az $Y=1$ (igen eset) bekövetkezési valószínűsége p , várható értéke $E(Y)=P(Y=1)=p$, varianciája $Var(Y)=P(1-P)$. A Cox féle logit transzformáció (Cox, 1964) alkalmazása után (valószínűségi hányadost logaritmálása és erre lineáris egyenlet illesztése):

$$\log it(p) = \beta_0 + \beta_1 x = \log\left(\frac{p}{1-p}\right), \text{ ahol az odds a } \frac{p}{1-p} \text{ esélyhányados.}$$

Továbbra is kérdés az, hogy milyen eséllyel valósulhat meg az $Y=1$ és $Y=0$ eset. Ismert az Y eloszlása, ezért célszerű Maximum Likelihood becslési eljárást alkalmazni. A likelihood függvényként felírt bináris változók eloszlásának logaritmusát deriváljuk p szerint⁷, vagyis:

$$\frac{\partial(\sum_{i=1}^n y_i \ln p + \sum_{i=1}^n (1 - y_i) \ln(1 - p))}{\partial p} = \frac{\sum y_i}{p} - \frac{\sum(1 - y_i)}{1 - p}.$$

Ezt nullával egyenlővé téve azt kapjuk, hogy $\hat{p} = \frac{k}{n}$, ami a relatív gyakorisággal történő becslést adja meg. Miután bevontuk a logit modellbe a magyarázó változókat, hogy az esélyek logaritmusa lesz a magyarázó változók lineáris függvénye, amiből a becsült valószínűség:

$$\hat{p} = \frac{\frac{p}{1-p}}{1 + \frac{p}{1-p}} = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}},$$

ami abból adódott, hogy

$$odds = \left(\frac{p}{1-p}\right) = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p),$$

ami az újonnan felírt *odds* logaritmalásából adódott. Ez utóbbi egyenlet egyenlő $e^{\beta^T x}$. Ezt behelyettesítve regressziós paraméterek becsléséhez felírt Likelihood függvénybe⁸ kapjuk meg iteratív Newton-Raphson eljárással⁹, hogy

$$\log it(p_i) = \widehat{b}_0 + \widehat{b}_1 x_i,$$

ahol $\widehat{b}_1$ előjele határozza meg a logisztikus függvény alakját. Ebből kifejezve kapjuk meg, hogy

$$\hat{p} = \frac{1}{1 + e^{-(b_0 + b_1 x)}},$$

ami nulla és egy közé esik. Itt a b_1 előjele a logisztikus görbe alakját, még a b_1 a növekedés sebességét fejezi ki. (Hunyadi, 2004) (Adler, 2010)

A logisztikus regresszió eredeti formájában nem értelmezhető, a koefficiensek értelmezését úgy kapjuk meg, hogy e-adszorosát vesszük. Tehát az x_2 magyarázó változó nő, akkor átlagosan az *odds* (esélyhányados) e^{β_2} -szorosára változik, nő vagy csökken. Ha az e^{β_2}

⁷ A számlálóban már maga a likelihood függvény van felírva.

⁸ A Likelihood függvény kibővített alakja: $L(b) = \prod \left[\frac{\exp(\sum_j b_j x_{ij})}{1 + \exp(\sum_j b_j x_{ij})} \right]^{y_i} * \left[\frac{1}{1 + \exp(\sum_j b_j x_{ij})} \right]^{1-y_i}$

⁹ A Newton eljárás során a módszer olyan értékpárokat keres, ahol a Likelihood függvény a maximumát veszi fel.

kisebb mint egy, akkor az *odds* csökken. Ennek a belátásához helyettesítsünk be az alapösszefüggésbe:

$$\log \left[\frac{\frac{p(x+1)}{(1-p(x+1))}}{\frac{p(x)}{(1-p(x))}} \right],$$

ami a különbségvétel és néhány egyszerű átalakítás után egyenlő

$$(b_0 + b_1(x + 1)) - (b_0 + b_1x) = b_1.$$

Ez multiplikatív hatás. (Kovács, 2006) (Tétényi, 2013) A szolgáltató váltás ($Y=1$) feltételes valószínűségét pedig az oddsok hányadosaként kaphatjuk meg:

$$\frac{odds}{1+odds} = \hat{y}_i = \frac{e^{c+\beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{c+\beta_1 x_{i1} + \dots + \beta_p x_{ip}}},$$

tehát ez az érték annak a valószínűségét adja meg, hogy vált-e a fogyasztó bizonyos feltételek mellett. Ezeket rekordokra értelmezzük. Érdekes a kovariánsok szintjén ábrázolni a feltételes valószínűségeket, így látható egyedi eseteken keresztül a feltételes valószínűség alakulása. Ezt én is megteszem a gyakorlati fejezetben.

Egy *másik egyszerűbb származtatási módszert* mutatok be Fülöp Péter (2002) tanulmánya alapján, amely a látens változó vizsgálatán alapul. Az alábbi jelölések bevezetése szükséges: $y_n \in [1 \dots N]$, ahol az N a minta elemszáma; $\mathbf{x}_n$, az n -edik magyarázó változó k elemű oszlopvektora; β az együtthatók k elemű oszlopvektora, valamint az u_n a hibateg. Emiatt az y_n két kimenetű, jelen esetben is dichotom. A látens változós megközelítés szerint feltesszük, hogy y^* látens függő változó, ami az eredeti modell magyarázó változóit tartalmazó lineáris regresszióval is leírható:

$$y_n^* = \mathbf{X}_n' \beta + u_n, \quad (1)$$

ahol $E(u_n | \mathbf{X}_n) = 0$ és $Var(u_n | \mathbf{X}_n) = \sigma^2$ és

$$y_n = \begin{cases} 1, & \text{ha } y_n^* > 0 \\ 0, & \text{ha } y_n^* \leq 0 \end{cases} \quad (2)$$

Ha az u_n véletlen, és bevezetjük a G maradéktag eloszlásának függvényét, akkor (2) feltétel szerint:

$$\Pr(y_n = 1 | \mathbf{x}_n) = \Pr(y_n^* > 0 | \mathbf{x}_n) = \Pr(\mathbf{X}_n' \beta + u_n > 0 | \mathbf{x}_n) = G(\mathbf{X}_n' \beta).$$

A logisztikus eloszlás függvény *logithoz* kapcsolódó származtatása könnyen belátható, itt csak a módszer belátása miatt a lényegi pontokat közlöm a leírtakra alapozva Hunyadi László (2004) alapján.

$X \sim U(0,1)$ legyen a változónk, ekkor ha Y a változó logitja, akkor $Y = \log\left(\frac{x}{1-x}\right)$. Így az X sűrűség függvénye formálisan:

$$f_x(x) = 1, \text{ még } f_y(y) = f_x(g^{-1}(y)) * |J|, \quad (3)$$

a transzformáció következtében. Ezután a transzformált alakból fejezzük ki az X -et:

$$e^y = \frac{x}{1-x}, \quad x = \frac{e^y}{1+e^y} \quad (4)$$

majd a (4) egyenletet y szerint deriváljuk, így visszahelyettesítve megkapjuk az Y sűrűség függvényét¹⁰, ami nem más, mint a természetes (standard) logisztikus eloszlás sűrűségfüggvénye, amiből az eloszlás függvény:

$$F_y(y) = \frac{1}{1+e^{-y}}.$$

További paraméterek bevezetésével nem foglalkozok a dolgozat témája miatt. Bináris logisztikus modell esetén feltesszük, hogy a logisztikus eloszlás megegyezik a standard logisztikus eloszlással, amit nagy lambdával jelölünk. (Long, 1997) Ezért:

$$\Pr(y_n = 1 | \mathbf{x}_n) = \Lambda(\mathbf{X}'_n \beta) = \frac{1}{1+e^{-\mathbf{x}'_n \beta}}.$$

Fontos kiemelni még a *cut value* (vágási érték) fogalmát, amelynek értékét a modellépítő határozza meg. A modell ez alapján klasszifikálja a valószínűségeket. A vágási értéktől nagyobb értékeket az $Y=1$ kategóriába sorol a módszer, még az ettől kisebb értékeket az $Y=0$ kategóriába. A módszer tehát az együttes bekövetkezési valószínűség, azaz a likelihood függvény logaritmált alakjának a maximumát keresi, azaz a β becslése ML becslésen alapszik (Maximum Likelihood¹¹). Hajdu (2003) és Ben-Akiva et al. (1999) részletesebben kifejti a logisztikus regresszió és eloszlás származtatását, alkalmazását, ám összességében úgy vélem, hogy a módszer ilyen szintű tárgyalására szükség volt, azért, hogy

¹⁰ $f_y(y) = \frac{e^{-y}}{(1+e^{-y})^2}$, a sűrűség függvényt úgy kapjuk meg, hogy belátjuk, hogy a (3) egyenlet jobb oldalának első tagja egy, a második tagot $|J|$ úgy számoljuk ki, hogy a transzformált formából az X kifejezése után az y szerint deriváljuk a (4) egyenletet.

¹¹ További részletek a módszerről: Hunyadi et al. 2012b

kielégítő bizonyossággal használhassam a fogalmakat a gyakorlati fejezetekben, valamint hogy magabiztosan alkalmazhassam a módszert a gyakorlati probléma feltárására.

3.1.3. A modell illeszkedésének jóságát vizsgáló mutatók

„Ahogy Veall–Zimmermann megjegyzi, keresztmetszeti adatok esetén például 0,1 körüli R^2 érték mellett sem kell feltétlenül elvetni a modellünket, értékük ugyanis nagyban függ az adatok jellegétől.”¹²

A következőkben megadjuk a legfontosabb modell jóságát tesztelő módszereket, ugyanis egy modell tesztelés nélkül nem minősíthető. A dolgozat terjedelme miatt nem áll módomban az összes mutató részletes kifejtése. Mivel a modelleket ML-elv alapján becsüljük, ezért többnyire ML-becslésekre épülő specifikációs tesztek alkalmazunk a modellek tesztelésére. Mindegyik teszt egy nullhipotézissal és egy ellenhipotézissel rendelkezik, ahol általában a H_0 azt foglalja magában, hogy a magyarázó változók paramétereinek értékei nullák. A Lagrange multiplikátor teszt (LM) azt vizsgálja, hogy a loglikelihood függvény értéke különbözik-e nullától. Az Likelihood arány teszt (LR) esetén pedig a loglikelihood függvény értékeit vetem össze.

A magyarázó változókra vonatkozó tesztek esetén fő célunk a felesleges változók kiszűrése. Jelen esetben az ML becslés konzisztens¹³, de nem hatásos. A Wald teszt nullhipotézise kimondja, hogy az adott magyarázó változók nem szükségesek a modellbe, ami formálisan felírva, ha feltesszük, hogy $\beta_1 = \beta_2 = 0$:

$$H_0: Q\beta = r, \text{ vagyis } \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Így a H_1 -hez felírt összefüggés:

$$W = [Q\hat{\beta} - r]' [QVar(\hat{\beta})Q'] [Q\hat{\beta} - r], \text{ ahol } W \sim \chi^2.$$

¹² Fülöp Péter (2002) 272. oldal hivatkozik Veal-Zimmermann 1996-os tanulmányára: Pseudo R2 Measures for Some Common Limited Dependent Variable Models, amiben az írópáros kifejti, hogy amellet, hogy a 0,1-es érték nem ad okot a modell elvetésére, a 0,2 érték már egészen öröndetes lehet, viszont a 0,02-es érték már problémás. Más és más érték várható akár ágazattól függően is. Valamint felhívja a figyelmet a mutató számok eltérő értékei mögött áll módszertani különbségekre is: „For example, we shall discuss entirely typical cases with 1000 observations where one Pseudo-R, the McFadden R, will be about 0.25 while another, the McKelvey Zavoina R, will be about 0.5.” (Veal-Zimmermann, 1996, 4. oldal)

¹³ Konzisztencia és hatásosság fogalmáról bővebben: Hunyadi László (2009) Statisztikai következtetés elmélet című könyv 120.-130. és 88.-100. oldal.

Hiányzó változók tesztelése ritka, gyakorlatban ritkán fordul elő, így ezt nem fejtem ki.¹⁴ A globális tesztek közé tartozik az LR teszt, amelynek null hipotézisében az áll, hogy magyarázó változók paramétere nulla, azaz nincsenek hatással az eredmény változóra, ami az alábbi összefüggés segítségével számolható:

$$LR = -2(\ln L(c) - \ln L(\hat{\beta})),$$

ahol $L(c)$ és $L(\hat{\beta})$ a loglikelihood függvény értékét jelöli és $LR \sim \chi^2_{v(k-1)}$. A modell globális teszteléséhez a Pearson-féle χ^2 illeszkedés vizsgálatot végezhetjük el, ami a reziduumokon alapszik. A null hipotézisben a becsült modell helyessége áll. A reziduumot úgy kapjuk meg, hogy az eredményváltozó értékeiből (y_i) kivonjuk a becsült értékeket ($\hat{y}_i$), ahol

$$\hat{y}_i = \frac{e^{c + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{c + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}.$$

A sztenderdizált reziduumok ennek ismeretében:

$$r_i = \frac{(y_i - \hat{y}_i)}{\sqrt{\hat{y}_i(1 - \hat{y}_i)}}, \text{ ahol } \sum_{i=1}^n r_i^2 = X^2.$$

Az X^2 χ^2 eloszlást követ $n-(p+1)$ szabadságfokkal. A módszer hasonlít a lineáris esethez. Magas p érték esetén a null hipotézist elfogadjuk, vagyis a modell megfelelő. A χ^2 eloszláson alapuló Hosmer-Lemeshow teszt csak akkor kerül alkalmazásra, ha nem teljesül, hogy a becsült gyakoriságok nagysága legalább 5. A Lagrange-multiplikátor teszt a Pearson-féle χ^2 illeszkedési elven alapszik:

$$\chi^2 = \sum_x \frac{(f_x - n_x p_{xb})^2}{n_x p_{xb} (1 - p_{xb})},$$

ahol az f a megfigyelt és np a várt gyakoriságot jelöli. Használata széleskörű. A Hosmer-Lemeshow teszt akkor is használható, ha a kovariánsokhoz csak egy rekord tartozik. A módszer a becsült valószínűségek sorrendbe állításán alapszik, amiből g csoportot képzünk. Alapfeltétel, hogy egy csoportban nem lehet kevesebb, mint öt rekord. A részletes levezetést Hosmer-Lemeshow (1999) tartalmazza, aminek bemutatására nem térek ki, mivel nem képezi szorosan a dolgozat témáját. A loglikelihood függvényeken alapuló pszeudo R^2 mutatók közül a leginkább elterjedt a MacFadden-féle R^2 :

¹⁴ Bővebben: Maddala (1995)

$$R_{MF}^2 = 1 - \frac{\ln L(\hat{\beta}) - (k+1)}{\ln L(0)}.$$

A kezdő idézetben már utaltam a McKelvey Zavoina-féle pseudo- R^2 -re, amely egy más megközelítéssel a modell által magyarázott eltérés négyzetösszege alapszik:

$$R_{MZ}^2 = \frac{\sum_{i=1}^N (\hat{Y}_i^* - \bar{Y}^*)^2}{\sum_{i=1}^N (\hat{Y}_i^* - \bar{Y}^*)^2 + N\hat{\sigma}^2},$$

ahol $\hat{\sigma}^2 = \frac{\pi^2}{3}$, mivel a 2.2.2. alfejezetben bemutatottak alapján ez a sztenderd logisztikus eloszlás szórás négyzete. (Veall et al., 1996) (Fülöp, 2002) A modell „determinációs együtthatói” közé sorolható még a Cox-Snell és Nagelkerke R^2 is.¹⁵ Összességében elmondható, hogy a szakemberek között a mutatók rangsorolását tekintve nincs teljes egyetértés.

A grafikus alapú metódusok közül az elemzésemhez még a *ROC görbét* választottam, ugyanis a ROC (Receiver Operating Characteristics) görbe pont azt mutatja meg, hogy milyen vágási értéknél, mekkora arányban találja el a kategóriákat, azaz a bináris kimenetel illeszkedés jóságát vizsgálja.

		Valódi osztályok	
		p (elfogadjuk)	n (elutasítjuk)
A modell outputjai szerinti osztályok	Y (cut value fölött)	TP	FP
	N (cut value alatt)	FN	TN
	Összesen	P	N

3. ábra: Kontingencia tábla (Confusion Matrix)

Forrás: Fawcett, 2005 alapján saját szerkesztés

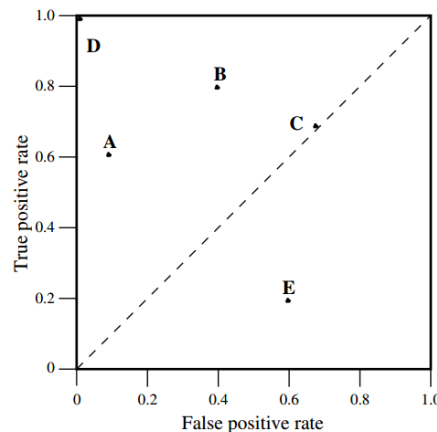
Bináris klasszifikáció esetén látható, hogy a TP (True Positive) akkor áll elő, ha a modell outputja p , még a valódi érték is p . Ha a valós érték n volt, akkor az érték FP (False Positive), ha az eredmény n és a valós érték is, akkor valós negatív (True Negative), végül, ha a valós érték p , még a modell szerinti besorolás n , akkor hamis negatív (False Negative) értékről beszélünk. A görbe ábrázolásához az alábbi főbb összefüggéseket kell számításba venni:

- $TP \text{ rate} \approx \frac{TP}{\Sigma P}$, ami az érzékenység (sensitivity)

¹⁵ A formális leírás bővebben: Kovács (2006), 155-157. o., ahol a látható, hogy a két mutató egymásból építkezik. A gyakorlat alapján már a 0,2 érték is kielégítő lehet.

- $FP\ rate \approx \frac{FP}{\sum n}$, ami a False Pozitive ráta
- $1-FP\ rate =$ specificitás (specifity)

A ROC görbe X tengelye a TP ráta, azaz az érzékenység, még az Y tengelyen az 1-speificitás, azaz az FP ráta látható. Így a ROC görbe a trade off-ot mutatja meg az haszon (TP) és költség (FP) közötti arálynak megfelelően.



4. ábra: 5 diszkrét klasszifikáció

Forrás: Fawcett, 2005

Látható, hogy (0,0) pont esetén a klasszifikáció nem követ el FP hibát, viszont nem is TP, aminek az ellentettje az (1,1) pont. Az ábrán a D pont (0,1) jelenti a tökéletes klasszifikációt, ahol a FP ráta nulla, még a TP ráta egy. Vagyis a ROC területen lévő pontok közül, relatíve annál jobb egy pont, minél „észak-nyugatabbra” helyezkedik el, tehát ahol a TP ráta magasabb, még az FP ráta alacsonyabb, tehát a B pont jobb, mint a C. A modell jószágáról a görbe alatti terület is árulkodik (ROC AUC, ROC Area Under the Curve), ami 0 és 1 közötti értéket vehet fel. Ha egy érték az átló alá megy, akkor a modell rosszabbul állapítja meg az értékeket, mint a véletlen, tehát a modell nagyon rossz. A 0,5 érték a véletlen besorolási képességnek felel meg. Minél nagyobb értéket kapunk, annál jobb a modellünk. A gyakorlatban a 0,7-es érték már jónak számít. Érdeemes megemlíteni még a ROC görbe AUC és a Gini index közötti elvi hasonlóságot: $Gini + 1 = 2 * AUC$ (Fawcett, 2005) ((S.n.), 2010). A modellem tesztelésére a teljesség igénye nélkül röviden bemutatott mutatókat is használni fogom, ezért volt szükség a modell ellenőrzési háttérének rövid bemutatása. Részletesebben a ROC görbéről többek között, és a görbe alatti területről Flach et al. (2003) és Fawcett (2001) tanulmányában olvashatunk, amik a görbe kiterjesztését, elméleti levezetését boncolják részletesen.

3.2. AZ R PROGRAM

„R kiemelekedően jó diagramok ábrázolásában, adatelemzésben, és modellek illeszkedés vizsgálatában.” (J. Adler, 2010, 17.o.)

Az előző fejezetben a módszertani elméleti háttérrel ismertettem meg az olvasót részletesebben. Mivel az elemzésem nagyméretű, automatizált és egy funkció akár több napig is futhat, ezért *mindenképp szükségesnek tartom* röviden bemutatni az R-t, mivel a dolgozatom magját a gyakorlati probléma programozása adja. Az R programozás segítségével érhető el a maximális információ kinyerése a mintából, ezért is fontos a gondolkodásmód, a megközelítés tömör ismertetése. Így ebben a fejezetben fő célom a módszertani háttér gyakorlati R programnyelvbeli alapok bemutatása, tehát az, hogy az elméleti felvezetés hogyan működik a valóságban a programkódok segítségével. Ezért az általam lefejtett néhány fontosabb programkód bemutatása a több száz oldalból elengedhetetlen, hogy az olvasó megértse, mi zajlik a háttérben, ugyanis a gyakorlati fejezetben már csak az eredmények közlésére fogok kitérni. A programkód a Mellékletben látható, ám az összes mellékszámítást a Melléklet korlátozott terjedelme miatt ott sem tudtam feltüntetni.

A dolgozatomhoz kapcsolódó adatelemzést tehát az R alapú R Studioban írom. Az R programnak nincs felhasználói felülete, vagyis parancssorok megadásával, vagy programfunkciók megírásával történik számolás. Az R program a nevét megalkotóiról kapta: Robert Gentleman, Ross Ihaka. A programválasztás ötletét a Mikroökonometria kurzus során elkészített R feladatok segítették, a Passaui Egyetem végzett kutatásom és az alábbi előnyös R tulajdonságok. Az IBM SPSS grafikus felületű, ám a többi menüpont eléréséhez syntax írása kell, ezért a teljes megoldást csak a kettős felhasználással kapható meg. Az R nem rendelkezik grafikus felülettel, ingyenes, nyílt forráskódú világszerte elterjedt programnyelv, statisztikusok, matematikusok, adatelemzők és adatbányászok használják. Operációs rendszer független alkalmazás, hardver és szoftver független alkalmazás, akárhol futtatható (Windex, Linux, Unix, Ubuntu). A világon az egyik legfejlettebb statisztikai programnyelv, kiváló grafika outputokkal. Összetett, több mint 4800 csomag él hozzá, amit állandóan fejleszt jelenleg egy 19 fős fejlesztő programozó csapat. Ezenkívül a bugokat bárki javíthatja.¹⁶ A csomagokat csak egyszer kell letölteni az `install.packages` paranccsal, utána elérhető a `library` paranccsal. Adatstruktúrája szerteágazó, valamint könnyen összekapcsolható más programokkal, kiterjesztésekkel, programnyelvekkel is (MS Excel, MS Access, MySQL, Oracle, LATEX, SAS, Pdf, CSV, C++, Java stb.).((S.n.), 2012a)

¹⁶ A bug a program nem várt viselkedését kiváltó programkód részlet, ami javításra szorulhat.

3.2.1. R objektum orientált programnyelv

Az R Studio fő részei a programkód (szkript), ahol a programkódot írja meg a felhasználó, a console, ahol az outputokat, az elvégzett számításokat írja ki a parancssorok alapján program, a táblázatok megjelenítő plot output, valamint az adatváltozó nézet ablak, ahol látható, hogy az elemző milyen változásokat hajtott végre a változói, adattábláin.

Az R program egy egyedi megoldás az objektum orientált programozásra. . Az R programozási nyelv nyelvi elemeit képezik az azonosítók, literálok, jelkészletek és kulcsszavak. A jelkészlet a latin ábécét és néhány egyéb karaktert foglal magában, a kulcsszavakkal kapcsolatos karaktersorozatok pedig nem használhatjuk azonosítóként (például if, else, NULL, TRUE, FALSE). A vezérlési szerkezet az értékadást, szekvenciát, elágazást és ciklust tartalmazza. Az R nyelvben függvények vannak, amihez saját argumentum lista és környezet tartozik. A függvény maga is objektum, ezért argumentumként és visszatérési értéként is használható. Ebben a nyelvben nincsenek sablonok. Bármely objektumnak lehet class attribútuma, azaz osztályszintű tulajdonsága. Ha egy objektum típusa dataframe, nem határozzuk meg alaphelyzetben a tulajdonságait, azaz implicit tulajdonságokkal bír, jellemező rá a lekérdezhetőség (például, hogy a táblánk X sorral és Y oszloppal rendelkezik). Viszont ha átalakítottuk, akkor már explicit tulajdonságokkal fog bírni. A generikus függvények használata praktikus, amit a további fejezetekben én is használok. A generikus függvény miután megállapította, hogy az argumentumként beillesztett objektum melyik osztályból származik, kiválasztja, majd meghívja a kívánt függvényt (`fun<-function(object, ...)`) Az első számú argumentum foglalja magában a meghívandó metódust. ((S.n.), 2013) Az imént felsorolt jellemzők gyakorlati alkalmazása a későbbi gyakorlati fejezetekben látható. A programkódok megértéséhez és magabiztos, körültekintő használatához a téma ismertetése elengedhetetlen.

3.2.2. Az R programnyelv gyakorlati alkalmazása a dolgozatban felállított hipotézisre

Ebben a részben a gyakorlati fejezethez szükséges főbb programkódjaimat foglalom össze, magyarázom.

A generikus funkciók megírásával alapoztam meg a modellem felépítését. Ez az alfejezet azért kiemelkedően fontos, mert tanúbizonyságot teszek egy általam megalkotott algoritmusról. Az algoritmus szót azért mertem használni, mert egy definiálható problémát oldok meg olyan eszközzel, amely ennek a problémának a megoldására eddig nem volt képes, továbbá más eszközök által használható funkciók sem fedték le az általam elvárt információkinyerési képességet, megoldási folyamatot, gyorsaságot és feldolgozási

képességet korlátok nélkül. Valamint ezeken a kódokon alapul a szakdolgozatom és ez jelent átkötést az elméleti és gyakorlati tudás között. A hipotézis vizsgálatom során a kezdetbeli intuitív modellek megalkotása után célom volt az adatbázisból kinyerhető összes információ megszerzése, ezért egy modellkeresési algoritmust programoztam le az egyedi funkcióim segítségével. Úgy vélem, hogy az IBM SPSS program segítségével nem lehet ilyen hatékonysággal programozni. Az R program nagyobb szabadságot, optimalizációs lehetőséget ad a program nyelve segítségével. Ez a dolgozat Magyarországon egyedinek számíthat, ugyanis gyakorlati jelenségek kutatása, feldolgozása hazánkban R segítségével a szakirodalmat kutatva még nem terjedt el, nem úgy, mint számos európai országban.

A funkcióknak a későbbi kutatásaim céljából, az egységesség és átláthatóság érdekében és a Passaui Egyetemen megszerzett R tudás alapján angol nevet adtam.

A programkódjaimat két fő csoportra bontottam a probléma megoldása érdekében. Az egyik csoport a funkciókat, még a másik csoport a gyakorlati alkalmazásokat, vagyis a funkciók felhívását, különböző táblákon való alkalmazását jelenti. Így a gyakorlati kódban elegendő egy hivatkozással a már korábban megírt funkciók behívása az egyszerűbb kezelhetőség érdekében. A gyakorlati probléma vizsgálatának módszertani hátterének megértése miatt is fontos a főbb egyedi funkciók megismertetése az olvasóval. A teljesség igénye nélkül csak a legfőbb egyedi script részleteket ragadom ki magyarázat céljából. A kód többi része a Mellékletben található. Az alapvető technikai részletekre a törzsszövegben kitérni nem célom, ezek szintén a Mellékletben láthatóak. Az egyedi funkciók megírása az automatizáció miatt fontos. Első lépésként az adatok beolvasására írtam egy funkciót, ami szintén elősegíti az automatizációt. A return parancs adja vissza az összes parancssorban a funkciót. A zárójelben a további függvény argumentumok láthatók, például, hogy legyen-e fejléc, valamint, hogy mi legyen a fájl neve.

```
> functions.readData <- function(fileName, sep, dec) {  
+   return (read.csv(file=fileName, header=TRUE, sep=sep, dec=dec))  
+ }
```

Az alábbi funkció az adatok kiírását könnyíti meg csv kiterjesztésű fájlba, valamint egyedi idő mintát tesz a fájl név végére a könnyebb kezelhetőség érdekében, amit a Sys.time() jelöl.

```
> functions.writeData <- function(frame, name.of.file){  
+   name <- paste(name.of.file, "_", format(Sys.time(), "%Y%m%d_%H%M%S"), ".csv",  
+   sep="")  
+   write.csv(frame, name, row.names=TRUE)
```

```
+ }
```

Az alábbi funkció azért fontos, mert ennek a segítségével generálok dummy változót, vagyis ezzel a paranccsal hozom létre a leendő modellem bináris kimenetelű eredmény változóit, amiről a későbbi fejezetekben lesz szó. Az ifelse parancs segítségével adom be a két kimenetel lehetőségeit felülírva ezzel az eredeti változó értékeit. Amikor új adattáblát (dataframe) hozok létre a megadott adatok alapján, akkor az oszlopnevek automatikus újranevezésére írt kódrészlet is kiemelkedő. Ennek segítségével az új oszlop formátuma az új név és a régi névből fog összeállni a nyomon követhetőség érdekében.

```
> functions.createDummy <- function(table, baseColumn, valuesForTrue){
+   table$Dummy <- ifelse(baseColumn %in% valuesForTrue, 1, 0)
+   table$Dummy[is.na(table$Dummy)] <- 0
+   return (table)
+ }
>
> functions.renameColumns <- function(frame){
+   for (i in c(1:length(colnames(frame)))){
+     +   colnames(frame)[i] <- sapply(strsplit(colnames(frame)[i], "\\."), `[,`,
+ 2)
+   }
+   return (frame)
+ }
```

A hipotézisem bizonyítása érdekében először célzott modell építési technikát alkalmaztam, majd ezután a lehetséges magyarázó változók kiszűrése után a logisztikus regresszióhoz írtam kereső funkciót. Az első részben öt magyarázó változót kerestem általam megoldott szignifikancia szinten. Az alábbi kódrészlet négyváltozós kombinációkat hív fel. Kiolvasható, hogy a funkció neve tehát az öt magyarázó változóból álló szignifikáns modell keresésére utal. A funkció lényege tehát, hogy általánosságban fogalmazza meg az elvégzendő módszertani feladatot, amelynek paramétereinek helyére később bármilyen értéket beírhatok, bármilyen környezetben, ezért egyedi és egyszerű a kód. Ez egy összetett funkció, hiszen könnyen észrevehető a CheckIfValuesArePresent funkció is, amely kibontva a Mellékeltben található. Lényege az, hogy kiszűri a többször előforduló modelleket, mivel képes észlelni, hogy az x1, x2, x3 az ugyanaz mint az x2, x3, x1 modell.

Az öt változó könnyedén átírható négy változóra, amit ez a funkció is jelöl. Ezt a kódot használtam fel az adatbázisomra, ahol szintén meg tudtam határozni a szignifikancia szintet is, így csak azon magyarázó változók kombinációin alapuló modelleket írtam ki az R-rel, amelyeknek minden magyarázó változója szignifikáns. Ez a kód igen számolás igényes,

ami azt jelenti, hogy a kód körülbelül 36 órán át futott végleges formájában. Az utolsó paramétereknél adtam meg a kívánt értékek nevét az átláthatóság érdekében.

```
> functions.findSygn4 <- function(columns, Significance){
+   ID <- 0
+   result <- data.frame()
+   for(i in c(1:length(columns))){
+     for(j in c(1:length(columns))){
+       for(k in c(1:length(columns))){
+         for(l in c(1:length(columns))){
+           logitCoef <- summary(glm(Dummy ~ unlist(columns[i]) +
+ unlist(columns[j]) + unlist(columns[k]) + unlist(columns[l]), data=columns,
+ family=binomial))$coef
+           if (length(logitCoef[logitCoef[,4] <= Significance, 4]) > 4){
+             names <- c("Constant", colnames(columns[i]), colnames(columns[j]),
+ colnames(columns[k]), colnames(columns[l]))
+             if(!functions.CheckIfValuesArePresent(where.to.check=result,
+ names.to.check=names)){
+               ID <- ID + 1
+               resultrows <- data.frame(rep(ID), names, logitCoef[logitCoef[,4]
+ <= Significance, 4])
+               colnames(resultrows)[1] <- "ID"
+               colnames(resultrows)[2] <- "valtozo"
+               colnames(resultrows)[3] <- "p-ertekek"
+               result <- rbind(result, resultrows)
+             }
+           }
+         }
+       }
+     }
+   }
+   return(result)
+ }
```

A gyakorlati probléma demográfiai gyökereire és a kutatásaimra támaszkodva felmerült néhány esetleges kardinális magyarázó változó rögzítése, amit a findSygnWith2FixedXs funkció megalkotásával tudtam elérni. Könnyen meghatározható, hogy hány magyarázó változót szeretnék rögzíteni. Kutatásom folyamán új funkciók segítségével próbáltam meg tökéletesíteni a modellkeresési algoritmusomat, ezért a kód a ROC görbét (rocauc) is visszaadja, valamint a görbe alatti terület értékét is, ami az előző fejezetben került részletesebb bemutatásra.

```
> functions.findSygnWith2FixedXs <- function(non.fixed.columns, fixed.columns,
+ Significance){
+   ID <- 0
+   result <- data.frame()
+   for(i in c(1:length(non.fixed.columns))){
+     for(j in c(1:length(non.fixed.columns))){
```

```

+   logitCoef <- summary(logit <- glm(Dummy ~ unlist(non.fixed.columns[i]) +
+ unlist(non.fixed.columns[j]) + unlist(fixed.columns[1]) +
+ unlist(fixed.columns[2]), data=non.fixed.columns, family=binomial))$coef
+   if (length(logitCoef[logitCoef[,4] <= Significance, 4]) > 4){
+     names <- c("Constant", colnames(non.fixed.columns[i]),
+ colnames(non.fixed.columns[j]), colnames(fixed.columns[1]),
+ colnames(fixed.columns[2]))
+     if(!functions.CheckIfValuesArePresent(where.to.check=result,
+ names.to.check=names)){
+       ID <- ID + 1
+       frame <-
+ functions.NAOmitAndSubsetFrame(data.frame(non.fixed.columns$Dummy,
+ non.fixed.columns[,i], non.fixed.columns[,j], fixed.columns[,1],
+ fixed.columns[,2]))
+       rocauc <- roc(frame[,1], fitted(logit))$auc
+       resultrows <- data.frame(rep(ID), names, logitCoef[logitCoef[,4] <=
+ Significance, 4], rep(rocauc))
+       colnames(resultrows)[1] <- "ID"
+       colnames(resultrows)[2] <- "valtozo"
+       colnames(resultrows)[3] <- "p-ertekek"
+       colnames(resultrows)[4] <- "ROC-AUC"
+       result <- rbind(result, resultrows)
+     }
+   }
+ }
+ }
+ }
+ return(result)
+ }

```

Ez a kódrészlet a modell jóságának ellenőrzésére szolgál általánosságban, tehát ezt is bármire fel lehet hívni a későbbiek során, ezért egyedi, ezért is van nagy jelentősége. A tesztekhez létrehozza a tökéletes modellt, és a becsült modell és a tökéletes modell közti távolságot vizsgálja likelihood arány teszt segítségével (lrtest).

```

> functions.TestLogitWithSaturatedModel <- function(dataframe, logit){
+   dataframe <- na.omit(dataframe)
+   fit <- glm(dataframe$Dummy~factor(1:length(dataframe$Dummy)), family=binomial)
+   return (lrtest(logit, fit))
+ }
> functions.DistanceFromNullModel <- function(logit){
+   return(1-pchisq(logit$null.deviance-logit$deviance, logit$df.null-
+ logit$df.residual))
+ }
> fix(functions.NAOmitAndSubsetFrame)

```

A végző értékeléshez a feltételes valószínűségek vizsgálata a kovariánsoknál elengedhetetlen. Az egyszerűség kedvéért írtam meg ezt a kódot, ami minden egyes kimenetelre kiszámolja a feltételes valószínűségeket a bináris logisztikus regressziós egyenleteken és a koefficienseken alapulva, ahogy az az előző fejezetben látható volt. Az értékelés grafikusán is látható a lakossági magatartás könnyebb megértése érdekében.

```

> functions.calculateAllScenario <- function(frame, logit){
+   colname2 <- colnames(frame)[2]
+   colname3 <- colnames(frame)[3]
+   colname4 <- colnames(frame)[4]
+   colname5 <- colnames(frame)[5]
+   y <- c(unique(frame[,1]))
+   x1 <- c(unique(frame[,2]))
+   x2 <- c(unique(frame[,3]))
+   x3 <- c(unique(frame[,4]))
+   x4 <- c(unique(frame[,5]))
+   coefs <- logit$coefficients
+   result <- data.frame()
+   for (j in x1){
+     for (k in x2){
+       for (l in x3){
+         for (m in x4){
+           col1 <- c(paste(colname2, ":", j, " ", colname3, ":", k, " ",
colname4, ":", l, " ", colname5, ":", m))
+           value <-
exp(coefs[1]+(j*coefs[2])+(k*coefs[3])+(l*coefs[4])+(m*coefs[5]))
+           col2 <- c(value/(1+value))
+           frame <- data.frame(col1, col2)
+           result <- rbind(result, frame)
+         }
+       }
+     }
+   }
+   return(result)
+ }

```

Az adatok előkészítése érdekében írtam meg ezt a funkciót. Könnyen belátható, hogy a kördiagramokhoz az adatok előkészítése érdekében kezeli a hiányzó értékeket, sorba rendezi az adatokat, és az adatoszlopból több dimenziós mátrixot generál, úgy, hogy minden különböző adattípusból vizsgálja, hogy mennyi érték van, és ezekhez hozza látra a mátrixokat, azaz csoportosítja az adatokat. A többi grafikus megoldás programozási háttere a Mellékletben látható.

```

> functions.groupArrayToMatrix <- function(array, transposed=FALSE,
valuesToCheck=NULL){
+   array[is.na(array)] <- "NA"
+   values <- unique(array)
+   if (!is.null(valuesToCheck)) {
+     for (v in valuesToCheck){
+       if (!(v %in% values)){
+         values <- c(values, v)
+       }
+     }
+   }
+   values <- sort(values, decreasing=FALSE, na.last=TRUE)
+   m <- matrix(nrow=length(values), ncol=2)
+   row <- 1
+   col <- 1
+   for(value in values){
+     m[row, col] <- value
+     m[row, col+1] <- length(array[array==value])
+   }
+ }

```

```

+   row <- row+1
+ }
+ if(transposed){
+   m <- t(m)
+ }
+ return(m)
+ }

```

Az előző funkciók segítségével a modellkeresés és vizsgálat után egy szűrőt állítottam be, aminek a segítségével kiírtam azokat a potenciális modelleket, ahol a p érték az általam megadott korlátokon belül van (ideális esetben egy százalék), majd ezeket exportáltam egy csv fájlba, ami szinten tartalmazza a jósági mutatókat is. A csv fájl Microsoft Excelbe importált alakja a Mellékletben látható. Ez lényegében egy összefoglalást ad a módszertan gyakorlati programozásáról valós adatokon. A kódom négy fontos paraméterrel dolgozik, ami alapvetően arra épül, hogy az előzőekben leírt szignifikancia kereső funkciónak a visszatérési értékét csv fájlba elmentve várja paraméterként. Általánosságban arra is használható, hogy általam kitalált véletlenszerű változókkal hozzon létre logit modellt, amennyiben birtokoljuk az adatsort. Fontos megemlíteni, hogy az elvárt result.frame (ami a paramtére) arra a sémára épüljön, hogy első oszlop tartalmazza az aktuális függvény egyedi azonosítóját (ID) a második oszlop pedig a változók neveit. A data.frame paraméter pedig a változók neveivel egyező adatábázist kell, hogy tartalmazzon. A result.frame-ben előforduló összes változó nevéhez tartozó adatsornak jelen kell lennie a data.frame-ben, mivel a függvényhez és további tesztekhez szükséges adatokat innen nyeri ki a kód. Miután az adott beolvasott változók neveihez párosítja a funkció az adatsorokat, megtisztítja azt majd a bináris logisztikus függvényt illeszt rá, valamint különböző jósági mutatók vizsgálatán alapuló tesztek futtatja le a kódom. Ezután az adott függvényhez elmenti a p értékeket, valamint a tesztek visszatérési értékét, így létrehozva egy táblázatot, amely tartalmazza az összes szükséges jósági mutató eredményt.

```

> functions.CreateLogitFromResultObjects <- function(data.frame, result.frame,
row.per.result, colnum.of.variable){
+   result <- data.frame()
+   ID <- 0
+   for(resultnr in c(1:(nrow(result.frame)/row.per.result))){
+     #create data frame for logit from the database with all datas, with the help
of result frame colnames
+     actualRowNr <- (resultnr*row.per.result)-(row.per.result-1)
+     y <- data.frame[,which(colnames(data.frame)=="Dummy")]
+     x1 <-
data.frame[,which(colnames(data.frame)==result.frame[actualRowNr+1,colnum.of.varia
ble])]

```

```

+   x2 <-
data.frame[,which(colnames(data.frame)==result.frame[actualRowNr+2,colnum.of.variable])]
+   x3 <-
data.frame[,which(colnames(data.frame)==result.frame[actualRowNr+3,colnum.of.variable])]
+   x4 <-
data.frame[,which(colnames(data.frame)==result.frame[actualRowNr+4,colnum.of.variable])]
+   frameForLogit <- na.omit(data.frame(y, x1, x2, x3, x4))
+   colnames(frameForLogit)[1] <- "Dummy"+   colnames(frameForLogit)[2] <-
as.character(result.frame[actualRowNr+1,colnum.of.variable])
+   colnames(frameForLogit)[3] <-
as.character(result.frame[actualRowNr+2,colnum.of.variable])
+   colnames(frameForLogit)[4] <-
as.character(result.frame[actualRowNr+3,colnum.of.variable])
+   colnames(frameForLogit)[5] <-
as.character(result.frame[actualRowNr+4,colnum.of.variable])
+
+   #Create logit, and run tests needed
+   logitCoef <- summary(logit <- glm(frameForLogit[,1] ~ frameForLogit[,2] +
frameForLogit[,3]+frameForLogit[,4]+frameForLogit[,5], family=binomial,
data=frameForLogit))$coef
+   names <- c("Constant", colnames(frameForLogit)[2],
colnames(frameForLogit)[3], colnames(frameForLogit)[4],
colnames(frameForLogit)[5])
+   ID <- ID + 1
+   #ROC
+   rocauc <- roc(frameForLogit[,1], fitted(logit))$auc
+   #Null Model
+   nullModel <- functions.DistanceFromNullModel(logit)
+   #saturate logit
+   saturateResult <-
summary(functions.TestLogitWithSaturaltModel(dataframe=frameForLogit,
logit=logit))
+   saturateResult <- c(saturateResult[4,1], saturateResult[5,1],
saturateResult[6,1], "", "")
+   resultrows <- data.frame(rep(ID), names, logitCoef[,4], rep(rocauc),
rep(nullModel), saturateResult)
+   colnames(resultrows)[1] <- "ID"
+   colnames(resultrows)[2] <- "valtozo"
+   colnames(resultrows)[3] <- "p-ertek"
+   colnames(resultrows)[4] <- "ROC-AUC"
+   colnames(resultrows)[5] <- "Null-model-distance"
+   colnames(resultrows)[6] <- "Saturate: Chisq, df, P"
+   result <- rbind(result, resultrows)}
+   return(result)}

```

Ebben a fejezetben a legfontosabb egyedi funkcióim programozását mutattam be az R programnyelv segítségével, ahol az alapvető technikai részletekre nem tértem ki. Célom volt megmutatni, hogy a kutatási kérdés feltárása érdekében az R programnyelv új távlatokat nyithat meg.

4. MODELL ÉPÍTÉS, TÉNYLEGES ELEMZÉS

A metódusok kellő ismeretével felvértezve, a megfelelő módszertani háttér megismeréséről bizonyosságot téve kezdem el a gyakorlati tapasztalatokon alapuló sejtések alapján a modell felépítését, a gyakorlati probléma vizsgálatát, azaz a hipotézisem tesztelését is.

„Alapvetően minden modell rossz, de néhány hasznos.” George Box

4.1. ADATOK

A hipotézisem bizonyítása céljából háromféle mintát használok fel a dolgozatomban, amelyből kettőnek az elkészítésében tevékenyen részt vettem, ami még jobban hozzásegített a probléma átlátásához.

4.1.1. Saját survey-en alapuló minta kisminta

Először egy kismintás survey alapján próbálom meg intuitíve sejtéseimet bizonyítani, amely saját egyedi kérdőívemen alapszik, amely a Mellékletben látható. Itt fő célom volt a jelenlegi kutatásom megalapozása gyenge keretek között. Hatvan embert kérdeztem le egyenlő arányban Borsod-Abaúj-Zemplén megyében és Budapesten. A kérdőív főbb pontjai az alapvető demográfia-változók mellett tartalmazzák a szolgáltató váltási hajlandóság kérdését, valamint a már egy korábbi munkámban kutatott kapcsolt áruszolgáltatások fontosságát a villamosenergia-szegmensben, az egyetemes szolgáltatásra jogosult fogyasztókat tekintve. Alapvetően a szolgáltató váltási hajlandóságot befolyásoló tényezőket, háttérváltozókat kutattam, valamint a két földrajzi terület közötti esetleges különbségeket a kapcsolat áruszolgáltatások megítélése és a szolgáltató váltási hajlandóság tekintetében. A kismintás survey minta természetesen nem reprezentatív, átsúlyozásra nem volt lehetőség.

4.1.2. Társadalomkutatási Intézet segítségével készült minta

Az imént említett kismintás kérdőívem középpontjában a szolgáltató váltási hajlandóság állt. Abban a lehetőségben volt szerencsém részesülni, hogy a TÁRKI Zrt. Adatfelvételi Osztályvezetője segítségével, András Zoltán úr segítségével, a kérdőívem egy része belekerült az országos Omnibusz kérdőívébe, ezért egy reprezentatív minta birtokosává válhattam, mivel részt vettem az elkészítési folyamatban. Az Omnibusz 2013/06 kérdőíve, amiben én is részt vettem és használtam a Mellékletben található. A TÁRKI Omnibusz egy havi rendszerességgel személyes lekérdezésen alapuló kérdőíven alapuló valószínűségi minta, amely a magyarországi lakosságot reprezentálja. A lekérdezésben csak 18 éven felüli és

állandó lakcímmel rendelkező személyek vehetnek részt. Az Omnibusz kutatás szlogenje a következő: "Döntések a társadalmi valóság beható döntésére.", amely jól összefoglalja az én kutatásom általános küldetését is. Egy Omnibusz lekérdezésnek több megrendelője van, hiszen a kérdőívben több téma is lekérdezésre kerül. A TÁRKI Omnibusz kutatások biztosítják a demográfiai háttérrel vizsgáló adatokat, amely kiemelkedően hasznos az alaphipotézisem bizonyítása céljából. Ilyen demográfiai adatok például családi állapot, iskolai végzettség, foglalkozás, család felépítése, jövedelem, vagyoni helyzet stb. Tehát 2013 júniusában a Tárki országosan lekérdezte saját kérdőívem egy részét, amely az imént Omnibusz energetikai részébe van beleépítve. Így egy országos átsúlyozott reprezentatív mintával rendelkezem, a további vizsgálataimat ezen végzem. A kismintás elemzésem segítséget nyújtott abban, hogy mit vizsgáljak a nagyminta esetén. Célom egy bináris logisztikus regresszió alapuló modell felépítése, amely képes arra, hogy négy magyarázó változóval magyarázza az eredményváltozó alakulásának feltételes valószínűségét, azaz, hogy a fogyasztó milyen valószínűséggel akar szolgáltatót váltani. Az előbbi esetet egyessel, még az utóbbit nullával kódolom. Úgy sejtem, hogy különböző háttérváltozók vizsgálatával lehet megragadni a váltási esélyt.

Az adatbázis az alkalmazott valószínűségi mintavételi eljárás miatt valószínűségi minta, tehát minden felnőtt magyar lakosnak azonos esélye van a mintába való bekerülésre. Ezáltal a mintából levont következtetések a teljes populációra azonosíthatóak. (TÁRKI, 2013ab) A minta elkészítése többlépcsős, arányosan rétegzett, valószínűségi mintavételi eljárás alapján. A rétegzés első lépcsőjénél minden megyéből kiválasztásra kerülnek azok a települések, amelyek a mintába kerülnek. Ezután került meghatározásra az egyes községekben élők száma a mintában. Az alapelv, hogy minden megye minden településtípusából akkora arányban kerültek a mintába lekérdezettek, amekkora az alapsokaság aránya a teljes népességben belül, megyére és megyén belüli településtípus lakosságállományra lebontva. Az kérdéseim adatfelvétele 80 településen és Budapest minden kerületében zajlott, összesen 207 kiindulási pont volt. A kérdezők véletlen sétás módszert alkalmaztak, ami egy bejárandó útvonalon és nem címeken alapul. A bejárandó útvonalon alapuló lekérdezés során kiválasztott háztartásokban Leslie Kish kulcsot alkalmaztak a kérdezők, hogy kiválasszák a megfelelő személyt az adott háztartásból.

Az adatbázisban négy dimenziós (nem -- korcsoport -- iskolai végzettség -- településtípus) egyéni súlyváltozó van, tehát a minta átsúlyozott, azért, hogy képes legyen reprezentálni a magyar lakosság adatait, azaz, hogy megfelelően illeszkedjen a felnőtt magyar

népességhez. A súlyozatlan és a súlyozott minta összehasonlítása az ismert demográfiai változókkal a Központi Statisztikai Hivatal eredményein alapszik. A Mellékletben található kérdőív releváns részeihez rendelkezem csak adattal, mivel az Omnibusz lekérdezés komplex kérdéscsoportokat felölelő egység. (TÁRKI, 2013ab)

4.1.3. Magyar Energetikai és Közmű-szabályozási Hivataltól felmérésén alapuló minta

Az utolsó adatbázis a már említett fogyasztói elégedettség vizsgálatához készült minta egy része a Magyar Energetikai és Közmű-szabályozási Hivatalból. A minta alapját véletlen jellegű kombinált mintavétel adja. Ennek a kérdőívnek¹⁷ az elkészítésében nem vettem részt. A kérdőív három fő részre oszlik. Tartalmazza a számomra fontos demográfiai változókat, valamint hálózati szolgáltatói tevékenységgel és ügyfélszolgálati tevékenységgel kapcsolatos változókat is. Ezt a mintát azért tartom érdemesnek elemezni, mert az egyik változó tartalmazza, hogy a megkérdezett lakossági fogyasztó milyen szolgáltatónál van, ami óriási jelentőségű információ, így a demográfiai változók mellé ezt is bele fogom építeni a modellbe.

4.2. KISMINTÁS SURVEY

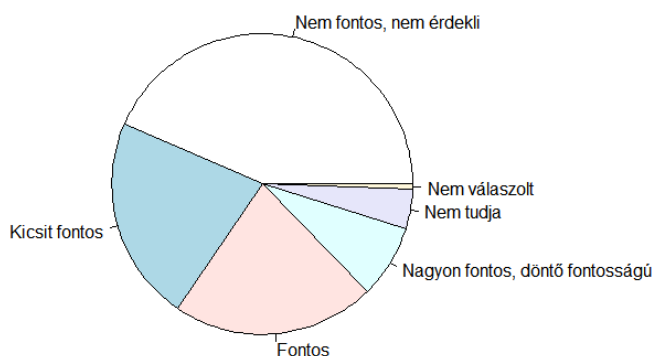
26 kérdésből álló kérdőívem segítségével 60 embert kérdeztem meg véletlenszerűen egyenlő arányban Budapestről és borsod-megyei kisvárosokból, községekből 2013 tavaszán a TÁRKI's felmérés előzményeként. Célom volt a budapesti és a vidéki lakossági fogyasztók közötti különbségek vizsgálata a váltás terén. A minta nem reprezentatív, átsúlyozás nem volt lehetséges, tehát a következtetések nem általánosíthatóak az egész populációra. Ám a budapestiek 28,3%-a még a Borsod megyeiek 71,7%-a gondolkozott szolgáltató váltáson, illetve a $p=0,39$ érték miatt a nullhipotézis nem elvethető, tehát nincs kapcsolat a két nominális ismérv (megye-váltás) között.¹⁸

Az alapszakos szakdolgozatom (2012. március) idején a Magyar Telekom verseny helyzetet teremtett az egyetemes szolgáltatásban, úgy hiszem, hogy a rezsicsökkentési hatás csökkenése után továbbra is veszélyhelyzetet fog jelenteni a Telekom az egyetemes szolgáltatóknak. Az egyetemes szolgáltatás ereje a tömegbázis, Magyarországon 5,1 millió lakossági háztartás tartozik jelenleg az egyetemes szolgáltatás alá. A versenyhelyzettel felrázott szolgáltatóknak a dolgozatomban azt a tanácsot adtam a szakmai gyakorlatom közben szerzett tapasztalatokra alapozva, hogy használják ki a tömegbázisuk erejét a többszektoros szolgáltatók ellen, és a fogyasztók megtartásáért kapcsolt áruszolgáltatásokkal

¹⁷ A kérdőív a Mellékletben található.

¹⁸ A számolás eredményei a Mellékletben láthatóak.

éljenek. Ez megvalósult a szakdolgozatom lezárása után. Ezért a kérdőívben vizsgálom a kapcsolt áruszolgáltatások fontosságát is. Céлом az intuitíve fölöslegesnek tűnő ajándék kuponok hatásainak vizsgálata. Úgy feltételeztem, hogy a borsod-megyei lakossági fogyasztóknak nem érdemes gyorséttermi kuponokat adni, hiszen drágább az eljutás az éttermekbe az eloszlásuk miatt, mint a levásárolható kedvezmény. Azt találtam, hogy a borsod megyeieknek nem fontosak a gyorséttermi kuponok. Továbbá megkérdejttem, hogy a jelenleg forgalomban lévő ajándékkuponok (gyors étterem, könyv stb.) jó döntésnek bizonyultak-e a lakossági fogyasztók megítélése alapján. Az Omnibusz nagyminta meglepő eredményt mutatott. A TÁRKI által rendelkezésemre bocsátott 2013/06 reprezentatív adatbázis alapján, amelynek az elkészítésében én is részt vettem, kiderül, hogy a fogyasztók több mint a felének fontosak a kapcsolt áruszolgáltatások, amely az 5. ábrán látható.



5. ábra: Fontos Önnek a kapcsolt áruszolgáltatás?

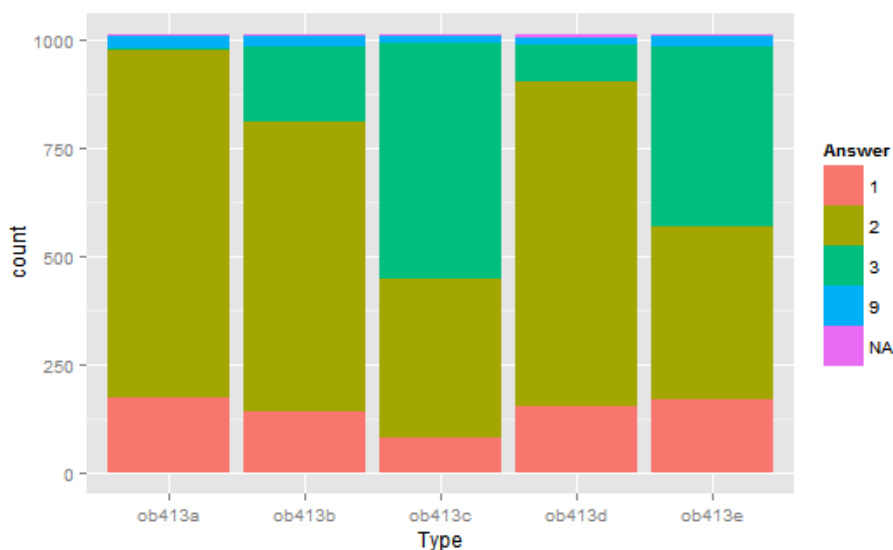
Forrás: saját kód alapján

Ebben a részben mélyebb értelmezésre a minta a kismintás átsúlyozatlan tulajdonságok miatt nem térek ki, kismintás survey-em csupán a sejtések megalapozására bizonyulhat hasznosnak. Mivel ezek nem csak a hatvan fős mintámon alapulnak így ezeket az eredményeket csak a Mellékletben közlöm. A modell építését a reprezentatív átsúlyozott mintán végzem a következő alfejezetben.

4.3. TÁRKI ADATBÁZISON ALAPULÓ MODELL

A modell építését az alapvető, a már módszertani fejezetben feltüntetett statisztikák elvégzése előzte meg. A TÁRKI adatbázisban szereplő változók neveit meghagytam a könnyebb nyomon követhetőség, ellenőrizhetőség és a későbbi kutatások érdekében, ugyanis

az eredeti táblában így szerepel. Valamint a Mellékletben található programkódok megértése miatt is szükséges az eredeti változónevek feltüntetése és a pontos dokumentáció. Célom a váltási hajlandóság háttérváltozóinak vizsgálata. A modellben a célváltozó az ob413a (VÁLT) lesz, azaz váltana-e a fogyasztó áramszolgáltatót?, amire a válasz alaphelyzetben igen (1) nem (2) NT (9) nincs ilyen szolgáltatása (3) NA. Az alábbi ábrán öt szolgáltatással kapcsolatban elhangzott hasonló kérdésre írtam ki egy oszlopdiagramot, ahol az ob413b (GÁZ), az ob413c a vezetékes telefon (VTELEFON), az ob413d (MOBIL), a mobil telefonszolgáltató még az ob413e (BIZT). a kötelező gépjármű felelősség biztosítás



6. ábra: Váltana szolgáltatót?

Forrás: saját kód alapján

Látszik, hogy az 1012 rekordból a célváltozó esetén, azaz a VÁLT a (3), (9) és NA kategória elenyésző, még a többi változó esetén ez nem mondható el, aminek a modellépítésnél lesz jelentősége.

Hipotézisem tehát, hogy különböző demográfiai háttérváltozók hatással vannak a szolgáltató váltási szándékra, ilyenek a kor, iskola, háztartás jövedelme, egy főre jutó jövedelem, háztartásban élők száma, így a vállalatoknak erre is tanácsos fókuszálniuk. Az árat azért nem építem bele magyarázó változóként a modellbe, mivel kevesebb, mint 1 Ft-os kWh-nkénti különbség van szolgáltatóként a regulált egyetemes szolgáltatói villanyár miatt. A rezsicsökkentés hatásai minden egyetemes szolgáltatót egyenlő arányban érintették. A dolgozatnak nem célja a rezsicsökkentés hatásainak elemzése, ez további nagy lélegzetvételű kutatásokat kíván, amit a terjedelmi korlátok miatt ezen kutatás keretei között nincs

lehetőségem bemutatni. Viszont itt felhívnám arra a figyelmet, hogy a lakossági fogyasztók nagyobb mértékben reagálnak a háztartási fogyasztók az áremelésre, mint az árcsökkentésre. (REKK,2007) és a becsült árrugalmasságuk -0,36 Sugár András által vezetett kutatás alapján, ami azt jelenti, hogy az ár 1%-os emelkedésének hatására átlagosan a lakossági körben 0,35%-kal csökkenti a fogyasztás mértékét. Ezért elmondható, hogy a lakosság árérzékenysége alacsony, azaz a villamos energia kereslet rugalmatlan, mivel a lakossági árrugalmasság értéke messze egy százalék alatt marad.

Az alapstatisztikák után függetlenség vizsgálatot végeztem a modell lehetséges magyarázó változóinak a függő változóval való kapcsolata feltérképezésére, ami a khi négyzet eloszláson alapszik. Az főbb eredményeket az alábbi R output részletben foglaltam össze. A második sorban látható a már átkódolt Y (Dummy) változó és az egyéb változók kapcsolata, valamint a létező esetek száma és a vizsgálathoz tartozó p érték.

```
summary(freqTable7 <- xtabs(~kor+Dummy, data=TARKI)) #X=A KÉRDEZETT KORA
#Call: xtabs(formula = ~kor + Dummy, data = TARKI)
#Number of cases in table: 977
#Test for independence of all factors:
#      Chisq = 71.35, df = 70, p-value = 0.4326
> summary(freqTable8 <- xtabs(~Dummy+netto, data=TARKI)) #x=NETTÓ HAVI KERESLET
Call: xtabs(formula = ~Dummy + netto, data = TARKI)
Number of cases in table: 971
Test for independence of all factors:
      Chisq = 182.79, df = 155, p-value = 0.06296
> summary(freqTable9 <- xtabs(~Dummy+iskola, data=TARKI)) #x=ISKOLAI VÉGZETTSÉG
Call: xtabs(formula = ~Dummy + iskola, data = TARKI)
Number of cases in table: 977
Test for independence of all factors:
      Chisq = 6.512, df = 8, p-value = 0.59
> summary(freqTable10 <- xtabs(~Dummy+neme, data=TARKI)) # x=A KÉRDEZETT NEME
Call: xtabs(formula = ~Dummy + neme, data = TARKI)
Number of cases in table: 977
Test for independence of all factors:
      Chisq = 0.10358, df = 1, p-value = 0.7476
> summary(freqTable11 <- xtabs(~Dummy+egyutt, data=TARKI)) #X=HÁZTARTÁS/FŐ
Call: xtabs(formula = ~Dummy + egyutt, data = TARKI)
Number of cases in table: 977
Test for independence of all factors:
      Chisq = 13.402, df = 9, p-value = 0.10452
> summary(freqTable11 <- xtabs(~Dummy+csalall, data=TARKI)) #X=CSALÁDI ÁLLAPOT
Call: xtabs(formula = ~Dummy + csalall, data = TARKI)
Number of cases in table: 977
```

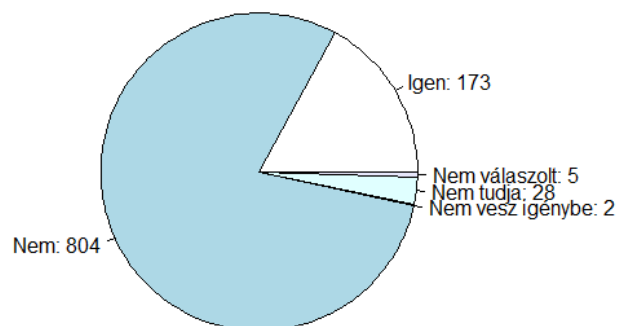
```

Test for independence of all factors:
      Chisq = 11.238, df = 5, p-value = 0.04686
> summary(freqTable11 <- xtabs(~Dummy+dolg, data=TARKI)) #X=MILYEN BEOSZTÁSÚ
Call: xtabs(formula = ~Dummy + dolg, data = TARKI)
Number of cases in table: 977
Test for independence of all factors:
      Chisq = 16.734, df = 8, p-value = 0.033
> summary(freqTable13 <- xtabs(~Dummy+lakter, data=TARKI)) #X=LAKÁS ALAPTERÜLETE
Call: xtabs(formula = ~Dummy + lakter, data = TARKI)
Number of cases in table: 974
Test for independence of all factors:
      Chisq = 143.46, df = 93, p-value = 0.0006136

```

A Dummy változó tehát egy már átkódolt változó, a logisztikus regresszió függő változója. A modellépítés során az adatbázisban az Y változó a „Ha egyszerűen és könnyen válthatna szolgáltatót, váltana Ön áramszolgáltatót?” kapott négy kimenetelű válasz: igen (1), nem (2), nem vesz igénybe (3), nem tudja (9), hiányzó adat (NA, -6) volt.

Váltana áram szolgáltatót ha könnyen tehetné meg?



7. ábra: Átkódolás előtt az Dummy_ob413a Y

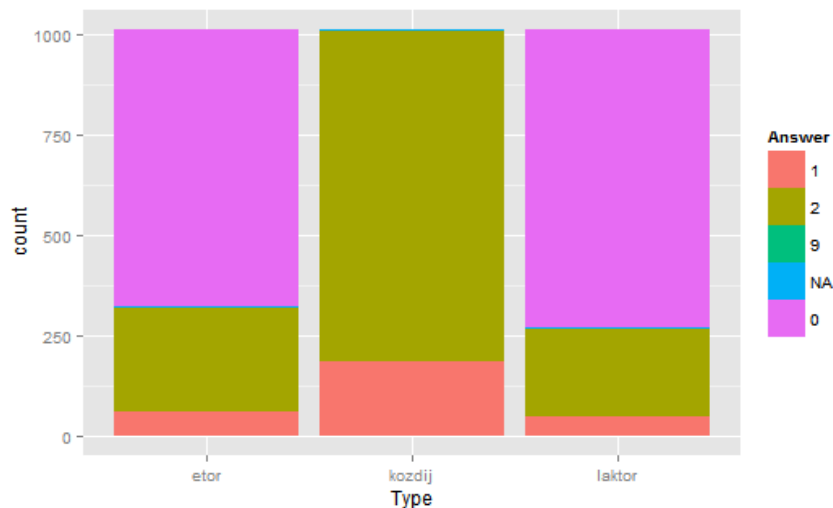
Forrás: saját programkód

Ezért a multinomiális logisztikus regressziós eszköztár alkalmazását nem tartottam megfelelőnek, hiszen nem célom a „nem vesz igénybe ilyen szolgáltatást”, „nem tudom” és system missing adatok jelenlétének magyarázása. Emiatt az ebbe a három kategóriába tartozó fogyasztókat töröltem, dummyt hoztam létre a fent látható funkció megírásával, így tehát látható, hogy a beavatkozás után 977 rekord maradt az 1012-ből. A hipotézisem első részét elvetem, vagyis nincs kapcsolat az alapvető demográfiai változók (kor, nem) és az eredmény változó között, mivel a magas p érték miatt a null hipotézis nem vethető el, vagyis a vizsgált

nominális ismérvek alakulása egymástól független. A kismintás survey alapján itt is vizsgáltam az egy háztartásban élő személyek (együtt) és az Y függetlenségét, ahol a p érték alapján a null hipotézist elvettem 10%-on, még a családi állapotot (csalall) 5%-on nőtlen, házas, élettárs, özvegy és elvált kimenettel. A megye változó hasonlóan szignifikáns lehet, ahol a p érték majdnem nulla. Az „Ön jelenleg dolgozik/nyugdíjas?” kérdésre adott válasz és az Y változó között van kapcsolat, ahogy erre már Rekettye Gábor is rávilágított. (Hetesi et al, 2001) Ám a modellépítésnél a leprogramozott algoritmus alapján nem bizonyult szignifikánsnak a változó. A létrehozott korváltozó és az Y között viszont nincs kapcsolat. Végül a lakás alapterülete (lakter) és az Y (Dummy) között a függetlenség elvethető $p=0,000003$ értékkel.

A függetlenség vizsgálat és az alapstatisztikák után a modellt építem fel a 2. alfejezetek alapján. Ahogy már bemutattam több funkciót dolgoztam ki a szignifikáns magyarázó változók keresésére, ha a szignifikancia szintet 5%-nak választottam meg. A megadott változók alapján bináris logisztikus regressziókat futtat, minden felírható kombinációban. Eredményül pedig azokat a függvényeket adja, amelyek minden változója az általam megválasztott szignifikancia szinten szignifikáns. A szignifikancia szinteket először 5 majd 1%-on írtam ki. A programkódot a közérthetőség érdekében áttördeltem, amely szintén a mellékletben látható.

A sejtésem alapján szóba jövő magyarázó változókat átkódoltam, ahol indokolható volt ez a lépés. A KOZDIJ változó arra a kérdésre ad választ, hogy „Előfordult az elmúlt félévben, hogy pénzühiány miatt nem tudták befizetni a közüzemi díjat?”, ahol igen(1), nem(2), nem tudja (9), nincs ilyen költség (0) és a system missing. Az egyéb tartozásos tételeket is ábrázoltam is vizsgáltam a kapcsolatot a célváltozóval a tisztítás előtt. Ennek eredménye látható az alábbi ábrán.



8. ábra: Előfordult-e már, hogy pénzühiány miatt nem tudták befizetni az alábbi tételeket?

Forrás: saját programkód alapján

Az ETOR változó az egyéb hiteltörlesztést, a LAKTOR pedig a lakáshiteltörlesztést mutatja. Mivel a másik két változónál sok rekord esetén nincs ilyen költség, ezért ezzel a továbbiakban nem foglalkozok. A második oszlopban látható, hogy a 9, NA és 0 kategória elhanyagolható, nem okoz sérülést, így a gyakorisági tábla alapján a két fogyasztót töröltem, hiszen csak két egyed esik bele a harmadik és ötödik kategóriába. Mivel a demográfiai változók nem voltak szignifikánsak a modellben, így ezeket elhagytam. Az alábbi modell adódott:

```
call:
glm(formula = Dummy ~ ob413b + ob413c + kozdij + egyutt, family = binomial,
    data = tisztitottTabla1)
```

Coefficients:

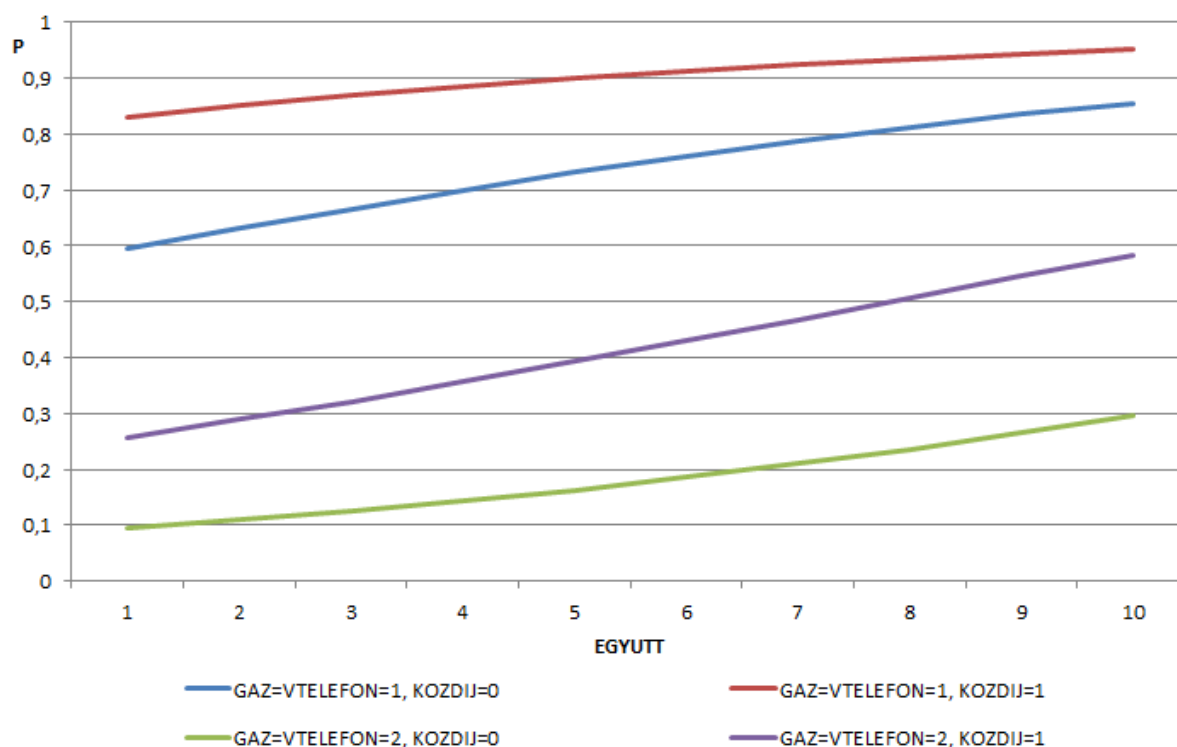
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.86503	0.47358	6.050	1.45e-09 ***
GÁZ_ob413b	-2.29314	0.20036	-11.445	< 2e-16 ***
VTELEFON_ob413c	-0.34683	0.14793	-2.345	0.0190 *
KOZDIJ_kozdij	1.20224	0.23594	5.095	3.48e-07 ***
EGYUTT_egyutt	0.15525	0.07268	2.136	0.0327 *

Number of Fisher Scoring iterations: 5

A modell magyarázó változói hatással vannak az eredményváltozó feltételes valószínűségének alakulására, ahol a Dummy a kétkimeneteles kódolt eredményváltozó Y, a GÁZ „Ha egyszerűen és könnyen válthatna szolgáltatót, váltana Ön gázszolgáltatót?”¹⁹ A kimenetek az ob413b és ob413c kérdésnél: igen(1), nem(2) nem vesz igénybe (3), nem tudja(9). Mivel itt a vezetékes telefon esetén 300 rekord eltűnne, ha subset-tel csak az egyes és kettes kimenetelt hagyom, akkor 677 rekord maradna, így ezt nem teszem meg. A Dummy kódolás miatt a GAZ esetén a 9 érték kiesik. , VTELEFON ugyanez a kérdés, csak vezetékes telefonszolgáltatóra, az együtt, hogy hányan élnek egy háztartásban és a közdíj, hogy van a közmű tartozása vagy sem. A modell neve legyen **Logit1**. Az egyéb főbb eseteket szintén a Mellékletben közlöm.

Tehát a közdíj tartozások 3,32-szeresére növelik az esélyt, azaz az odds-ot. Több különböző kovariánst mutatok be: ha a fogyasztó nem akar se gáz se telefon szolgáltatót váltani és nincs közmű tartozása sem valamint a háztartásban három ember él, akkor 12% az esélye, hogy szolgáltatót váltana, viszont ha akar váltani gáz és telefon szolgáltatót is váltani és nincs tartozása, valamint öten élnek együtt, akkor a feltételes valószínűség 0,74. Viszont ha nincs vezetékes telefonja, de váltana gázt és van tartozása valamint egyedül él, akkor $P(Y=1)=0,87$. Ha nem váltana gáz szolgáltatót, nincs vezetékes telefonja, de van tartozása és öten élnek együtt akkor 58% az esélye, hogy vált, viszont, ugyanezekkel a feltételekkel ha nincs tartozása, akkor a váltás valószínűsége 12%-ra csökken. A különböző szolgáltatások együtt hatásából a többszektoros szolgáltatók is profitálhatnak. Mivel így nem teljesen átlátható a folyamat az összehasonlítás kedvéért a főbb eseteket egy grafikonon foglalom össze. A létező kovariánsok feltételes valószínűségét egy funkció leprogramozásával számoltam ki, így a valószínűségek gyökereik mentén könnyedén összehasonlíthatóak, amely a Mellékletben részletesen lekövethető. A kívánt kovariáns megfelelő értékeit behelyettesítem az egyenletbe, az e-adszorosát veszem, majd így megkapom az odds értékét, amit aztán leosztok az 1+odds értékével, amivel megkapjuk a p értékét. Ezt szerepeltetem a 9. ábrán.

¹⁹ Tehát kimenetek az ob413b_GÁZ és ob413c_VTELEFON kérdésnél: igen(1), nem(2) nem vesz igénybe (3), nem tudja(9). Mivel itt a vezetékes telefon esetén 300 rekord eltűnne, ha subset-tel csak az egyes és kettes kimenetelt hagyom, akkor 677 rekord maradna, így ezt nem teszem meg. A Dummy kódolás miatt az ob413b esetén a 9 érték kiesik. AZért neveztem el dummynak az eredmény változót, mert kétkimenetele van.



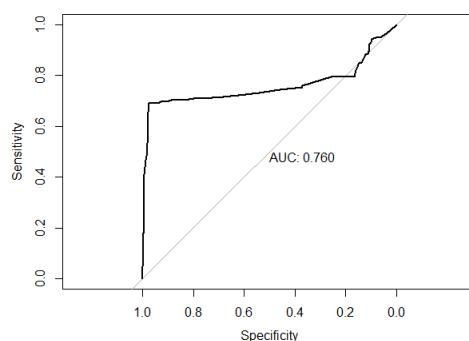
9. ábra: Feltételes valószínűségek megjelenítése, Logit1 modell

Forrás: saját programkód alapján

A 9. ábrán tehát csak néhány érdekesebb kovariánst hívtam fel. A feltételes valószínűségek tehát csak a meghatározott leszűrt esetekre vonatkoznak, azaz, ha az összes paraméter értéket kap. Látszik, ha se telefont se gáz szolgáltatót nem akar váltani a fogyasztó, akkor nagyobb a feltételes valószínűsége a váltásnak, ha van közműdíj tartozása, ami az egy háztartásban élő fogyasztók számának növekedésével szintén nő. Ha viszont a fogyasztó akar telefon és gáz szolgáltatót is váltani és tízen élnek egy háztartásban, valamint van közműdíj tartozása, akkor a váltás valószínűsége $P(Y=1)=0,93$. Azaz a váltás feltételes valószínűsége abban az esetben a legnagyobb a Logit1 modell szerint, ha gáz és vezetékes telefon szolgáltatás váltáson is gondolkodik a lakossági fogyasztó, van közdíj tartozása és a háztartásban minél többen, azaz tízen élnek együtt.

A modell jóságát is vizsgáltam. A Nagelkerke R Square 35%-os determináltságot jelez, ami a gyakorlatban kielégítő érték. Ezután függetlenség és illeszkedés vizsgálatot végeztem, azaz megnéztem, hogy a modellem mennyire távolodott el az üres modelltől és mennyire került közel a szaturált modellhez. A nullhipotézist a mellékletben látható alacsony p érték miatt elutasítom, tehát az aktuális modell eltávolodott a nullmodelltől. A másik teszthez szükséges először egy szaturált tökéletes modellt építeni az R-ben, ami minden egyes

megfigyeléshez tartalmaz értékeket, ez a második sorban látható, majd likelihood arány teszttel (lrtest) hasonlítom össze a két modellt. Látható, hogy a szaturált modell AIC értéke 1950, valamint a residual deviance nulla. Tehát jó az aktuális modellem. A vágási érték (cut) defaultja 0,5 volt, ami a klasszifikációs táblánál is fontos. Ilyen vágási érték mellett a modell 91,7%-os biztonsággal meg tudja becsülni az Y kimenetelét. Ám 0,6-os érték mellett ez tovább javul 92,6%-ra, ami azt jelenti, hogy közel 70%-os eséllyel helyesen jósolja meg a modell az Y=1 kimenetet, még közel 98%-os valószínűséggel pedig az Y=0 kimenetet. A ROC görbe azt mutatja meg, hogy milyen vágási értéknél, mekkora arányban találja el a kategóriákat, azaz a bináris kimenetelű illeszkedés jóságát vizsgálja. A görbe érdekes alakja a két kimenet megítélésének eltérő jóságából adódik. A tisztított táblán ez az érték 0,76 lett. A modell az előző tesztek kiállta, viszont a ROC görbe alatti viszonylag magas érték ellenére a modell nem minden kimenet esetén erős. Ezért célom a modell tovább fejlesztése, ami az alábbi fejezetekben látható.



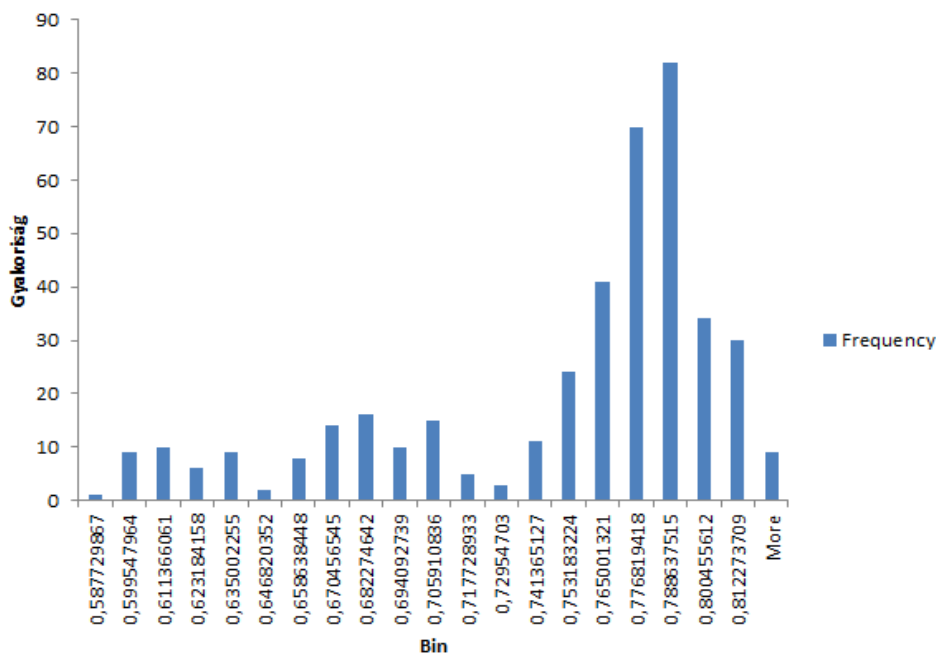
10. ábra: Logit1 modell ROC görbéje és a görbe alatti terület

Forrás: saját programkód alapján

4.4. ÚJ MODELLEK FEJLESZTÉSE AZ OPTIMALIZÁLT ALGORITMUS SEGÍTSÉGÉVEL

A Logit1 modell ROC görbéje szerint nem minden kimenetet becsült jól a modell, azaz nem kiemelkedően erős. A leprogramozott algoritmusom alapján, ami a 2.1.2. fejezetben is bemutatásra került, valamint az egyéb részletei a Mellékletben található, több százféle lehetőség adódott eltérő jósági mutatókkal, miután a kódom több napig futott a komplex számolási háttér miatt. A legalacsonyabb görbe alatti terület (Area Under the Curve) AUC 0,67 lett, még a legmagasabb 0,82. A hipotézisem teljes körű bizonyítása érdekében tanúbizonyságot kell arról tennem a dolgozatban, hogy felvértézve a módszertani és programozási ismeretekkel képes vagyok választani a modellek közül. Így miután a releváns magyarázó változókat kiválasztottam, az algoritmusom alapján számolt modelleket kell

értékelnem. Különbséget kell tenni a közel azonos AUC értékű modellek között is. A különböző modellek AUC értékeiből hisztogramot készítettem, hogy a megoszlásuk követ-e valamilyen eloszlást. Az Y tengelyen a gyakoriságot, még az X tengelyen a különböző görbe alatti területeket láthatjuk. Három csoport különíthető el, ami egyedinek számít, a jobb oldali csoportba tartoznak a legjobb modellek. Balra elnyúl az eloszlásuk, nincsen kimagaslóan jó modell, hanem a jobb értékek köré csoportosulnak a jobb modellek. Ezért további vizsgálódásra van szükség, hogy megtudjuk, hogy egyes változók mennyit adhatnak hozzá a modell jóságához, külön lebontva.

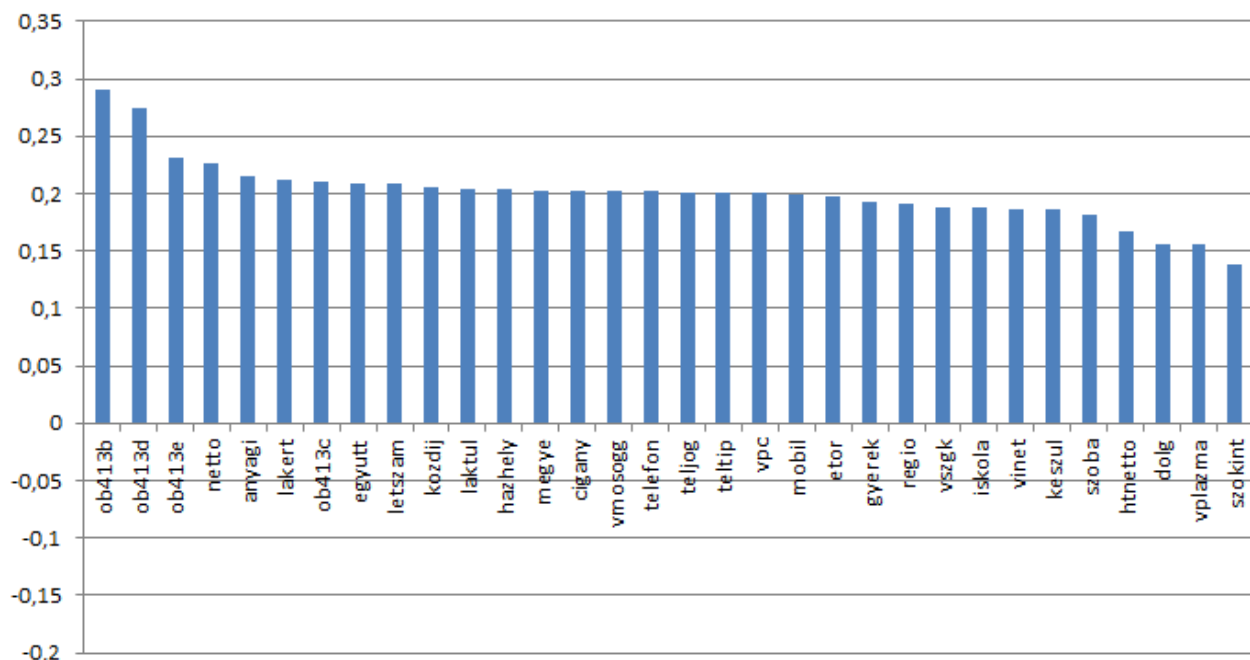


11. ábra: A feltételes valószínűségek ábrázolása hisztogramon

Forrás: saját kód alapján

Így azt vizsgáltam, hogy melyek azok a változók, amik a jobb modellekben szerepelnek csak, és melyek azok, amelyek a rosszabb modellekben így a végén le lehet szűkíteni arra, hogy melyek azok a változók, akik jobban teljesítenek az AUC-khez képest az átlagnál. Figyelembe kell venni, hogy ha három változó jól magyaráz, akkor az a negyedik magyarázó változót is felhúzhatja, tehát érdemes lebontani, hogy egyes változók mennyit tesznek hozzá az adott modellekhez. Minden releváns változóhoz egy számot rendeltem, egy vagy nulla, hogy melyik modellben szerepel, majd kiszámoltam a becsült AUC-t és a tényleges AUC-t és a négyzetes eltérést véve minimalizáltam ezt a solver függvénnyel. Így megkaptam a határ hozzájárulását az egyes változóknak, amit aztán csökkenő sorrendbe állítottam. Tehát így intuitíve a változók jóságát kaptam meg a modellekhez képest. Könnyen

látható, hogy a GÁZ változó a legjobb majd azt követi a váltana-e mobil szolgáltatót majd kötelező gépjármű felelősség biztosítást, majd ezután következik a havi nettó bevétel, az anyagi helyzet megítélése majd egyéb családdal kapcsolatos háttérváltozók. Az alapvető demográfiai változók nem szerepelnek.²⁰



12. ábra: A változók határhozzájárulása

Forrás: saját programkód alapján további számítások a határ hozzájárulásról

Látható, hogy az, hogy a kérdezett dolgozó-e, tanuló, vagy nyugdíjas szintén keveset az hozzá a modellhez, még az átlagos értékeit tekintve is. Vagyis az alapvető hipotézisem, miszerint az alapvető demográfiai változók befolyásolják a szolgáltatás váltás tekintetében a lakossági fogyasztót elvethető, viszont az egyéb távolabbi háttérváltozók között a megye és az iskola szerepel, ami nem elhanyagolható.

Ezen tapasztalatokkal felvértezve kezdem meg az új szofisztikáltabb modell felépítését. Így a hozzájárulási rangsor alapján tettem változókat az algoritmusba, amelynek

²⁰ A változók sorban: váltana e gáz/mobiltelefon/kötelező gépjármű felelősség biztosítást, havi nettó bevétele a háztartásnak, anyagi helyzet, ingatlan értéke, váltana-e vezetékes szolgáltatót, hányan élnek egy háztartásban a kérdezettől kívül, a kérdezettel, tulajdonos-e vagy bérlő a kérdezett, melyik megyében lakik, roma származású-e, mennyi mosógépe van, van-e vezetékes telefonja, milyen jogú településen él, mennyi PC-je van, van-e mobilja, van-e hiteltörlesztése, mennyi gyerek van, melyik régióból jött, mennyi személygépkocsija van, milyen iskolázottságú, szokott-e internetezni, mennyi az egész háztartás nettó bevétele, dolgozó-e, hány plazma televíziója van, végül milyen gyakran internetezik. Ezek a változók a mellékletben található kérdőívben pontosan megtalálhatóak.

eredményeként két modellt emelek ki.²¹ A kódban látszik a már tárgyalt funkció felhívása. A kutatási kérdésem szempontjából a megye változót emeltem ki, ami az alapvető demográfiai változók egyike. A koefficiensek az alábbi részletben láthatóak, amely az egyenlet alapját adja meg.

```
> tisztitottTabla4 <-
functions::NAomitAndSubsetFrame(data.frame(TARKI$Dummy, TARKI$ob413b,
TARKI$ob413d, TARKI$anyagi, TARKI$megye))
> summary(logit4 <- glm(Dummy ~ ob413b + ob413d + anyagi + megye,
data=tisztitottTabla4, family=binomial))
```

Call:
glm(formula = Dummy ~ ob413b + ob413d + anyagi + megye, family = binomial,
data = tisztitottTabla4)

Coefficients:

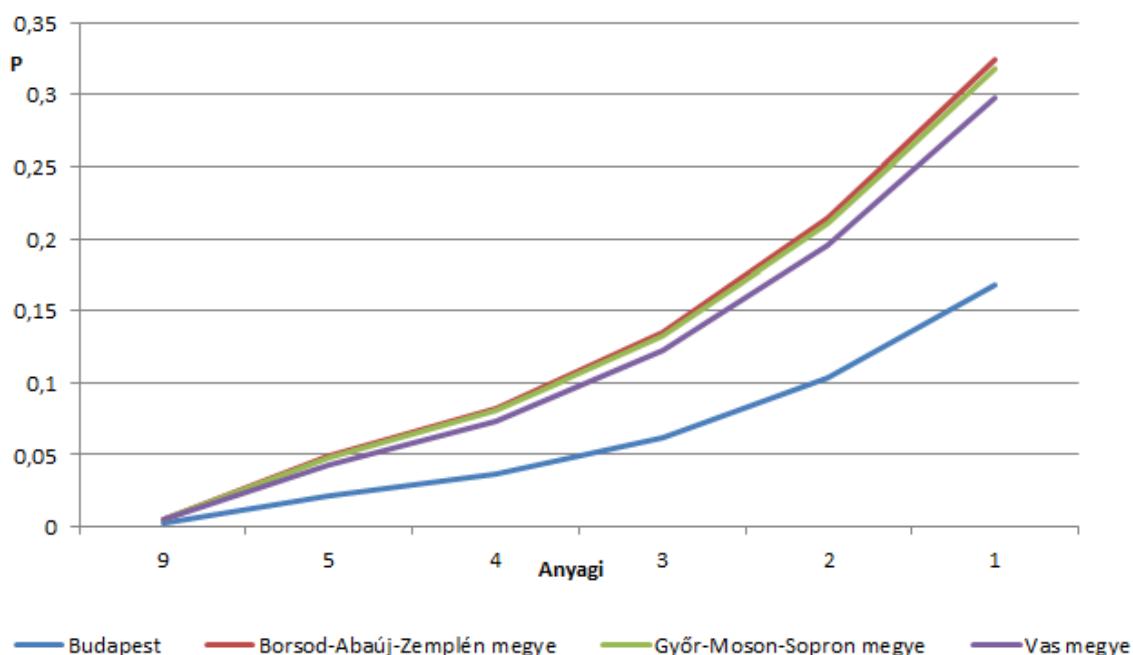
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	7.156846	0.713273	10.034	< 2e-16	***
GÁZ_ob413b	-2.047723	0.200879	-10.194	< 2e-16	***
MOBIL_ob413d	-1.596878	0.224660	-7.108	1.18e-12	***
anyagi	-0.558157	0.122282	-4.565	5.01e-06	***
megye	-0.009158	0.003277	-2.795	0.0052	**

A második modellben, aminek a neve legyen **Logit2** modell tehát négy változó szerepel, a „váltana-e gáz szolgáltatót” (GÁZ) a „váltana-e mobil telefon szolgáltatót” (MOBIL) a „milyennek ítéli meg az anyagi helyzetét”²² (ANYAGI) és a megye (MEGYE) változó. A koefficiensek önmagában nem adnak pontos iránymutatást a problémáról. Látható azonban, hogy csakúgy mint a **Logit1** modellnél, itt is negatív előjellel szerepelnek az „ob” változók, azaz, hogy akarna e váltani bármilyen egyéb releváns szolgáltatót a lakossági fogyasztó.

A feltételes valószínűségek kiírása segít megérteni a váltási hajlandóságot befolyásoló egyéb tényezőket, így az alábbi ábrán a kovariánsokat mutatom meg. A feltételes valószínűségeket úgy kaptam meg, hogy funkciót írtam az R-ben a folyamat gyorsítására, azaz, hogy visszaadja az összes kimeneti párt, visszahelyettesítse a bináris logisztikus egyenletbe, ebből megkapja az oddst minden kimenetre, majd ezt leosztja az (1+odds) összeggel, így a p előáll. Ezt használtam a Logit1 modellnél is. A feltételes valószínűségek lekérdezésének szűrése esetén minden kombinációhoz egy értéket kapunk, így előáll az alábbi ábra, amely feltételezi, hogy a GÁZ=MOBIL=2 vagyis nem akar váltani a fogyasztó gáz és vezetékes telefon szolgáltatót.

²¹ A többi modell a Mellékletben látható.

²² Ez a változó eddig részletesen nem szerepelt így ennek kimeneteit nem csak a Mellékletben, hanem itt is célom szerepeltetni: „Hogy érzi, Önök anyagiilag (5) gond nélkül él, (4) beosztással jól kijönnek, (3) éppen kijönnek, (2) hónapról hónapra anyagi gondjaik vannak, (1) nélkülöznek, (9) nem tudja.

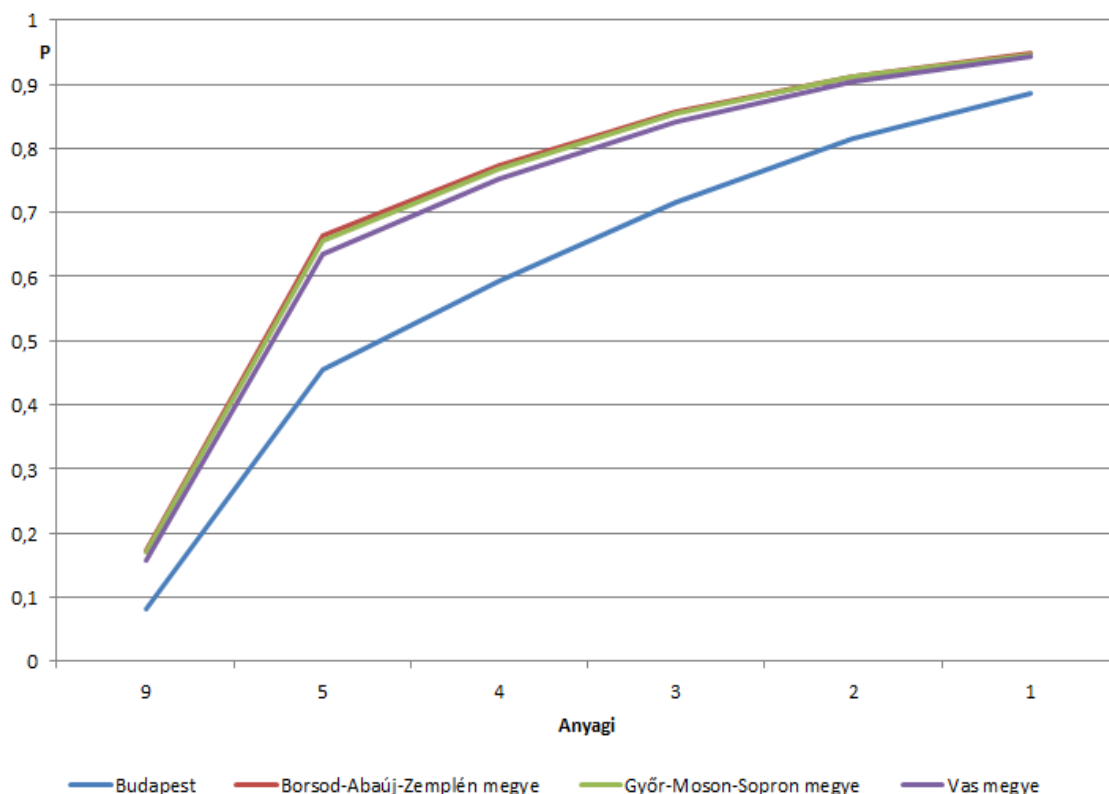


13. ábra: Feltételes valószínűségek megjelenítése, Logit2 modell

Forrás: saját programkód alapján

Az grafikonon a legeltérőbb helyzetben lévő megyéket jelenítettem meg összehasonlításképp. Az X tengelyen az anyagi helyzet szerepel, ahol az 5 a jó anyagi helyzetben lévő lakossági fogyasztót, még az 1 a nagy nélkülözések között élő fogyasztót jelöli, a 9 kategória pedig nem válaszolt a kérdésre. Az a borsod megyei lakossági fogyasztó, aki sem gáz sem mobil szolgáltató váltáson nem gondolkozik és jó anyagi körülmények között él, annak a szolgáltató váltási hajlandósága alacsony. Tehát itt is bizonyítást nyer a hipotézisem, miszerint a demográfiai változók hatással vannak a váltási hajlandóságra az egyetemes szolgáltatásra jogosult lakossági fogyasztókat tekintve. Összehasonlítva a borsod megyei kovariánsokat reprezentáló görbét az ábrán megjelentett budapesti lakosokkal látszik, hogy egy budapesti jó körülmények között élő lakossági fogyasztó nagyobb eséllyel gondolkodik a szolgáltató váltáson, mint egy borsod megyei. Az anyagi helyzet romlásával a váltási hajlandóság mind a négy közigazgatási területen nő, és a kiemelten nehéz anyagi körülmények között élő lakossági fogyasztók esetén éri el a csúcsát a borsodi régióban. A három megye közel azonosan mozog. Összességében elmondható, hogy a feltételes valószínűsége annak, hogy az a fogyasztó aki nehéz anyagi körülmények között él és sem mobil sem gázszolgáltató nem akar várni, nagyobb eséllyel gondolkodik a szolgáltató váltáson, mint a jó anyagi körülmények között élő. Itt felhívnom a figyelmet arra, hogy az olvasónak érdemes összevetni ezt az ábrát a **Logit1** feltételes valószínűségeivel a közdíj

tartozások megléte értelmében. Most nézzük meg mi történik, hogy mobilszolgáltató és a gázszolgáltató váltáson is gondolkodik a lakossági fogyasztó.

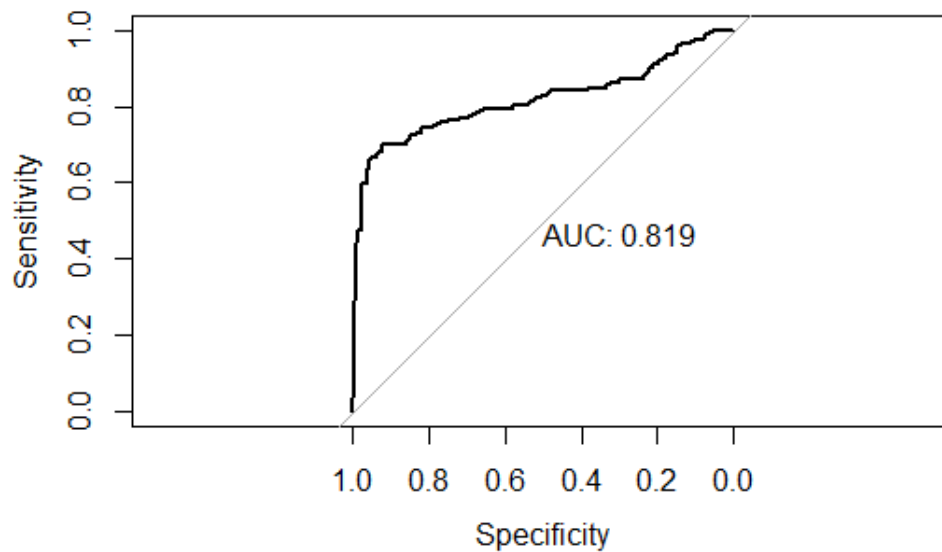


14. ábra: Feltételes valószínűség megjelenítése 2, Logit2 modell

Forrás: saját kód alapján

A P értékek itt 0,9 fölé mennek a rossz anyagi körülmények között élők esetén, tehát az anyagi körülmények nagy befolyással bírnak a Logit2 modellben. Az összes eset megvizsgálására a dolgozat kötött terjedelme miatt nincs lehetőség, ezeket így csak a Mellékletben közlöm.

A modell jóságát vizsgálva a szaturált modelltől és a nullmodelltől való távolságra vonatkozó megfelelő hipotézis is kielégül, vagyis a modell, aminek a neve legyen Logit2 modell jó. Viszont érdemes megnézni a bináris modellek illeszkedését vizsgáló ROC görbét. A ROC görbe és az AUC, azaz a görbe alatti terület kielégítő, sokkal jobb, mint az Logit1 modellnél, vagyis a célokat a modell fejlesztésével és a megfelelő algoritmus kifejlesztésével elértem.



15. ábra: ROC görbe és a görbe alatti terület, Logit2 modell

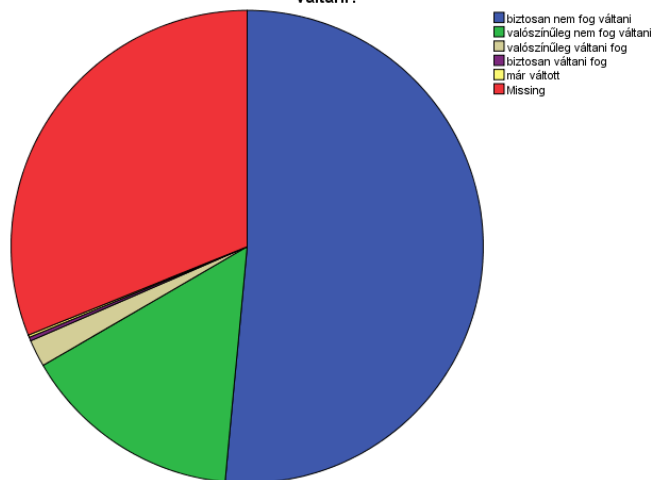
Forrás: saját kód alapján

Látható, hogy a Logit1 modell ROC görbéjéhez képest a Logit2 modell sokkal jobban sorolja be a kimeneteleket, azaz összességében a kettes modell erősebben teljesít.

4.5. MEKH-ES ADATBÁZISON ALAPULÓ MODELL

Az előző modellek esetén különböző demográfiai változókat is vizsgáltam, ám nem volt lehetőségem figyelembe venni azt a változót, hogy melyik egyetemes szolgáltatóhoz tartozik a lakossági fogyasztó. A MEKH adatbázison alapuló modell segítségével lehetőségem nyílt arra, hogy az egyetemes szolgáltatók is egy magyarázó változóvá válhassanak. Ebben az adatbázisban a v129-asra kódolt változót használtam fel eredményváltozónak, amely arra a kérdésre ad választ, hogy „Ön mennyire tartja valószínűnek, hogy három hónapon belül kereskedőt fog váltani?” Az adatbázis 2009-es, tehát itt sincs értelme bevezetni a rezsicsökkentés kérdését. A válaszokat tekintve hat kimenet volt lehetséges: biztosan nem fog váltani (1), valószínűleg nem fog váltani (2), valószínűleg nem váltani fog (3), biztosan váltani fog (4), már váltott (5), nem tudja (6) vagy hiányzó adat. A lehetséges kimeneteleket számszerűsítve az alábbi ábrát kaptam:

129. Mennyire tartja valószínűnek, hogy három hónapon belül kereskedőt fog váltani?



16. ábra: Eredményváltozó kimeneteleinek megoszlása

Forrás:saját kód alapján

A piros szelet a hiányzó adatot mutatja, ami azért ilyen magas, mert a v128-as kérdés, miszerint „Hallott-e már Ön arról, hogy mindenki szabadon megválaszthatja a szolgáltatóját?” determinálja a fő változót, ugyanis a kördiagramon szereplő kérdésre csak azok a fogyasztók adtak választ, akik a v128-as kérdésre igennel válaszoltak. Ennek a változónak az alakulását is vizsgáltam, ami a Mellékletben található, ám érdemleges eredményre nem jutottam, nem találtam szignifikáns magyarázó változókat a demográfiai változókat tekintve. Nem célszerű több csoportot vizsgálni, ezért a multinomiális logisztikus regresszió alkalmazása elvethető. Így összevontam az egyes kettes és a hármas négyes csoportot, átkódolást hajtottam végre, a többit pedig kitöröltem a mintából, így 5238 rekord maradt meg. Így a v129-es (Y) változót kétmenetelűvé tettem, ahol az egyes eset tartalmazza a biztosan/valószínűleg váltani fog eshetőségeket, még a nullás eset a biztosan/valószínűleg nem fog váltani eseteket, amit aztán bináris logisztikus regresszióval vizsgáltam. A programkód megírása utána az alábbi modell adódott, amelynek a neve legyen **Logit3** modell.

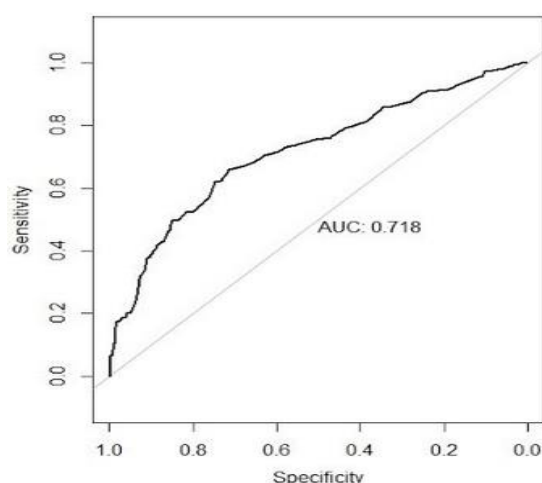
```
Call:
glm(formula = Dummy ~ v19 + v126 + d6 + terület, family = "binomial",
    data = MEH1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0855  -0.2688  -0.2022  -0.1491   3.3073

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.13282    0.69301   3.078 0.002087 **
```

v19	-0.57951	0.14569	-3.978	6.96e-05	***
v126	-0.74890	0.08204	-9.128	< 2e-16	***
d6	-0.61451	0.16697	-3.680	0.000233	***
terulet	0.22141	0.05106	4.336	1.45e-05	***

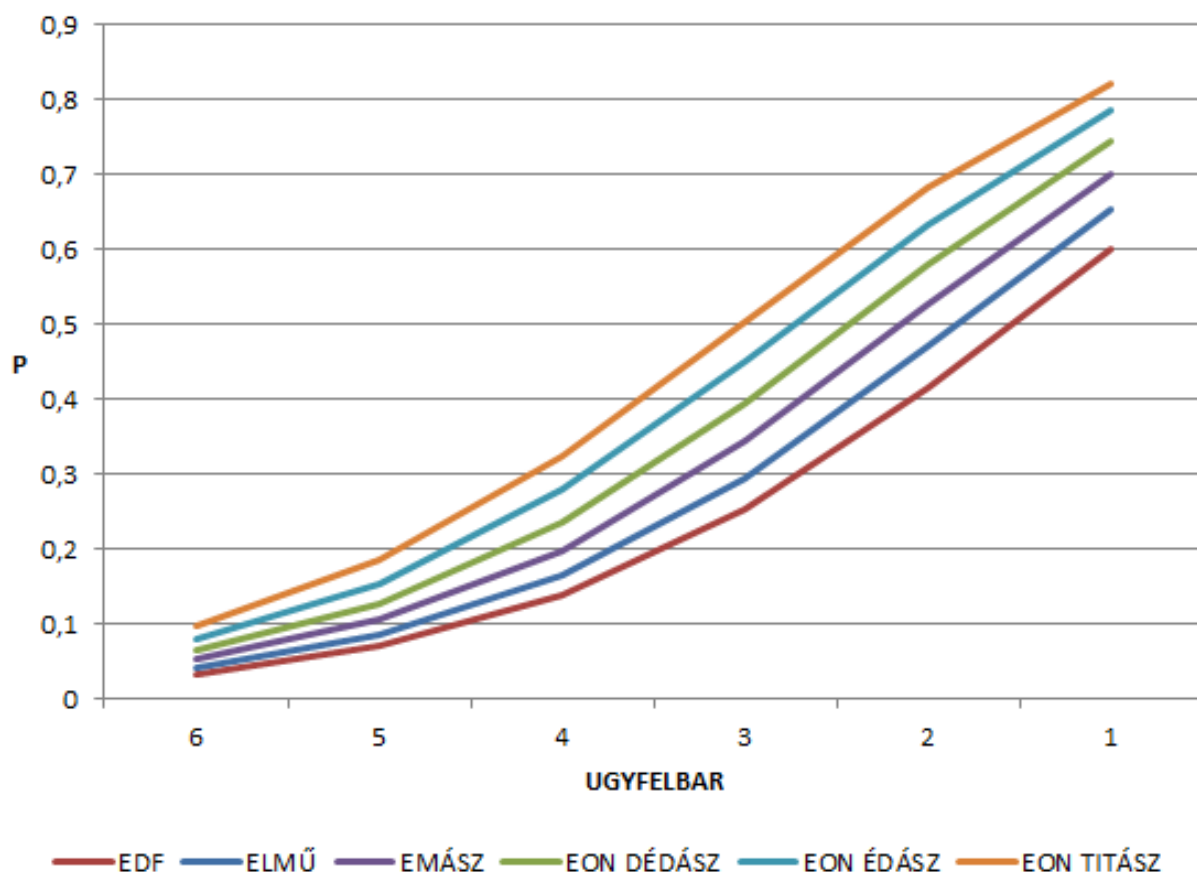
Tehát a Logit3 modell Y eredményváltozója az, hogy váltana-e a lakossági fogyasztó szolgáltatót, még a modellben lévő szignifikáns magyarázó változók a következők: „Mennyire fontos, hogy előre tájékoztassák a tervezett áramszünetekről?” (FONTOS), ami egytől négyig megy, ahol a négyes azt jelenti, hogy nagyon fontos; „Összességében hogyan minősítené az áramszolgáltatóját abban a tekintetben, hogy mennyire ügyfélbarát?” (UGYFELBAR), amelyre egytől hatig terjedő skálán kellett választ adni, ahol a hatos a nagyon jó; a NEM (férfi/nő) valamint a TERULET változó, azaz, hogy melyik egyetemes szolgáltatásban van a fogyasztó. Mind a négy magyarázó változó szignifikáns, a szaturált modellen alapuló hipotézis elfogadható, amely a Mellékletben látható. A kovariánsok értelmezése előtt azonban felhívnom a figyelmet arra, hogy a ROC görbe alatti terület nagysága alacsony, vagyis ez a modell a besorolási képessége gyengébb mint a Logit1 és Logit2 modell esetén, alacsony.



17. ábra: ROC görbe és görbe alatti terület, Logit3 modell

Forrás: saját kód alapján

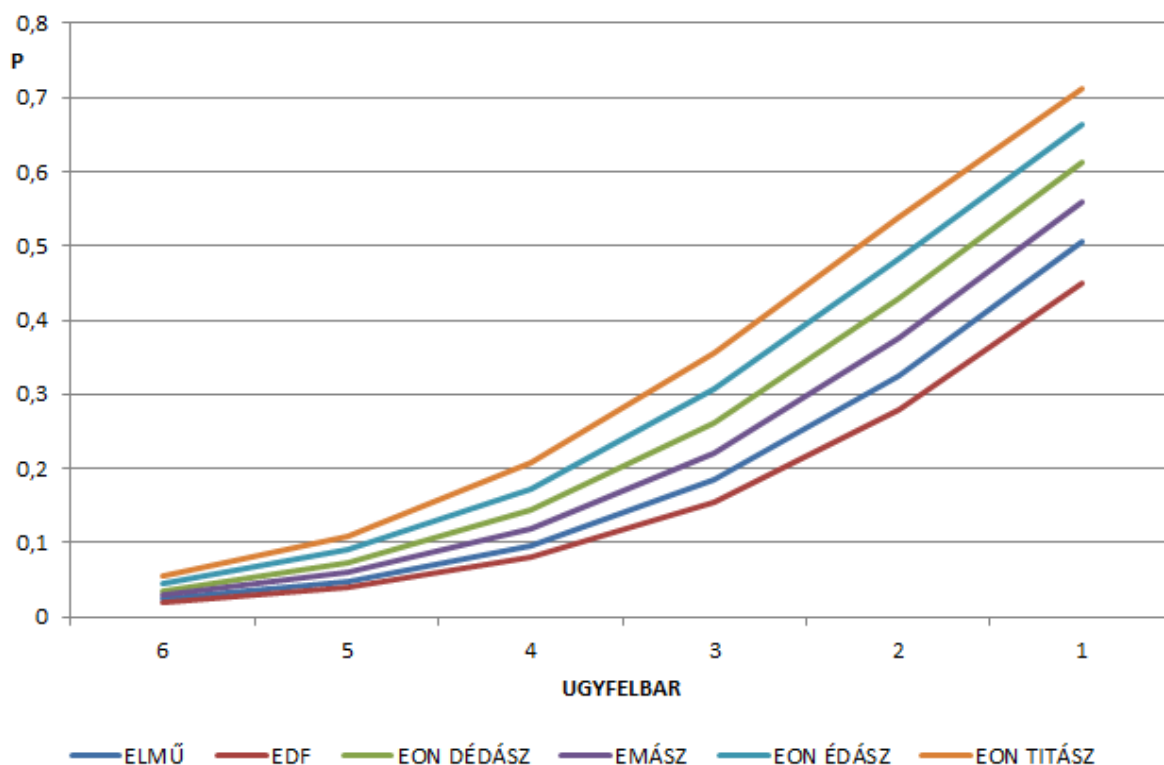
A feltételes valószínűségek szemléltetésénél a területbeli és a nemek közötti különbség bemutatására törekedtem. Először azokat a kovariánsokat írtam ki, amikor a lehetséges áramkimaradásokkal kapcsolatos tájékoztatások fontossága=1, vagyis ez nem fontos a lakossági fogyasztóknak, a nem változót pedig csak a férfiakra (nem_d6=1) szűrtem az X tengelyen pedig az ügyfélbarátsággal kapcsolatos kérdésekre adott válaszok láthatóak. A területbeli különbségeket emeltem ki.



18. ábra: A szolgáltató váltás kérdése a férfiak körében, Logit3 modell

Forrás: saját programkód alapján

Látható, hogy az a tény, hogy a fogyasztó mennyiben tartja ügyfélbarátnak a szolgáltatóját milyen mértékben hat a váltás feltételes valószínűségére. Tehát ha ügyfélbarátság szempontjából a lakossági fogyasztó rossznak (1) tartja a szolgáltatóját, akkor a váltás feltételes valószínűsége nő. Azonban felhívom a figyelmet arra, hogy ez az ábra a FONTOS változó miatt az ábra csak az egyes eseteket tartalmazza, azaz a fogyasztónak nem fontos, hogy előre időben tájékoztassák az áramkimaradásokról. Ugyanezt az esetet vizsgálom az alábbi ábrán is, viszont a nők esetében az egyetemes szolgáltatók szerint elkülönítve.



19. ábra: A szolgáltató váltás kérdése a nők körében, Logit3 modell

Forrás: saját programkód alapján

A nők esetében a maximális feltételes valószínűség 0,7 körüli azonos feltételek mellett. Számos egyéb kovariáns adata a Mellékletben található. Terjedelmi korlátok miatt nem tudtam kiemelni több lehetőséget.

Az imént elhangzott változók értékes eredményt jelenthetnek a vállalatok számára, így úgy vélem, hogy érdemes ezeket az összefüggéseket a későbbiekben egy nagyobb lélegzetvételű munkában tovább elemezni.

4.6. TANULSÁG, KÖVETKEZTETÉS

A célzott szolgáltató váltással kapcsolatos kérdéseimmel abba a szerencsés helyzetbe kerültem, hogy ezek országos szinten lekérdezésre kerültek a TÁRKI segítségével, ennek segítségével a kismintás survey-mel kapcsolatos kutatást tovább folytathattam országos reprezentatív mintákon. Összességében a három modell segítségével sikerült bizonyítanom a hipotézisemet a programozás segítségével.

A Logit1 modell esetén:

- szignifikáns magyarázó volt a GÁZ, VTELEFON, EGYUTT és KOZDIJ változó

- a szolgáltató váltás feltételes valószínűségét a villamosenergia-szegmensben az EGYUTT változó befolyásolja
- az EGYUTT változó növekedésével a feltételes valószínűség ($Y=1$) nő
- ha a lakossági fogyasztó vezetékes telefon és gáz szolgáltatót is akar váltani, akkor a villamosenergia-szolgáltató váltási esélye nő
- a maximumát pedig a 10 vagy több háztartás/fő számú háztartások esetén éri el
- a jósági mutatóknak megfelel
- vagyis családi háttérváltozók hatással vannak az eredményváltozó alakulására, eszerint a hipotézisem elfogadható.

A tökéletesített Logit2 modell esetén:

- egyéb demográfiai jellegű változókat is megkísértem a leprogramozott algoritmus segítségével belevonni
- szignifikáns változók: MEGYE, ANYAGI, GÁZ, MOBIL
- a jobb körülmények között élők kevesebb eséllyel gondolkodnak a váltáson
- a budapestiek kevesebb eséllyel gondolkodnak a szolgáltató váltáson
- a jósági mutatóknak megfelel
- azaz a hipotézisem itt is bizonyosságot nyert.

A Logit3 modell esetén:

- belekerült a nem változó a modell és a terület változó a modellbe, amelyeket segítségével értékes információkat nyertem ki
- vagyis a szignifikáns magyarázó változók: UGYFELBAR, FONTOS, NEM, TERULET
- a nők kevésbé hajlamosak a szolgáltató váltásra
- a szolgáltató ügyfelekhez való hozzáállása döntő fontosságú mindegyik egyetemes szolgáltató esetén
- a különböző egyetemes szolgáltatókhoz való tartozás ténye is hatással van a szolgáltató váltásra
- jósági mutatóknak kevésbé felel meg (AUC)
- a hipotézisem bizonyítást nyert

A programkódjaim hasonló jellegű problémák megoldására is alkalmas, hiszen könnyen örökíthető, a funkciók miatt felhasználható. A modellekkel sikerült bebizonyítanom,

hogy a szolgáltató vállalatoknak érdemes a demográfiai változókra is közelíteniük, valamint figyelembe venni azt is, hogy a más szolgáltatást is váltani szándékozók nagyobb eséllyel gondolkodnak a villamosenergia-szolgáltató váltáson is. Ez a komplex szolgáltatás csomagot nyújtó, azaz más néven többszektoros vállalatok stratégiája szempontjából fontos. Hiszen ha egy fogyasztó több szolgáltatás váltásán is gondolkodik, akkor a komplex szolgáltatás csomagokkal nagyobb eséllyel könnyebben meg lehet győzni.

Az egyetemes szolgáltatók monopol helyzete megtörhető a komplex szolgáltatáscsomagok megjelenésével, amit az első magyarázó változóm is alátámaszt, aminek segítségével a versenypiaci szolgáltatók képesek a regulált ár alá menni és emellett fejlesztéseket végezni. Az infrastruktúrák összekapcsolása lehet a jövő. Az szolgáltatóknak érdemes a közmű tartozást és a háztartásban lakó személyek számát is beleépíteni stratégiájukba. Az Megoldás lehet az is, ha az egyetemes szolgáltató versenypiaci leánya lefölözéssel csalogatja ki az egyszerű modell alapján előzetesen felmért váltani hajlandó fogyasztókat egy komplex szolgáltatás csomaggal. Kiterjesztett kutatásomban további céloom a tudatosság és elégedettség számszerűsítése, valamint a váltási hajlandóság és egyéb kereskedelmi és elosztói engedélyesi tevékenységek közötti kapcsolat felderítése (például áramkimaradást tapasztalt-e, ügyintézés gyorsasága, környezet védelem fontossága).

Tehát a hipotézisemet nem tudom elvetni. Valamint egyéb háttérváltozók jelenlétét és befolyását is felderítettem.

5. ÖSSZEGZÉS

Dolgozatomban az egyetemes szolgáltatásra jogosult lakossági fogyasztók magatartását elemeztem. Az első elméleti részben a kutatási kérdésem megalapozása érdekében alapvető definíciókat vezettem be, röviden összefoglaltam a lakossági fogyasztók helyzetét és az egyetemes szolgáltatás jogintézményét, majd megfogalmaztam a hipotézisemet, miszerint a magyar lakossági fogyasztók szolgáltató váltását demográfiai jellemzők is befolyásolják a villamosenergia-szegmensben. Megvilágítottam, hogy a kérdés nagy jelentőséggel bír, hiszen az okok egy része feltáratlan, valamint az elemzési eszközöket tekintve (R program) is egyedülálló a dolgozat. A kutatásomhoz motivációt a gyakorlati és kutatási tapasztalatom adta, valamint a Mikroökonometria tárgyam, ahol először találkoztam R kódokkal. Az R program kódokon alapul, tehát programozni szükséges.

A következő részben a felhasznált statisztikai eszközök kellő ismeretéről tett tanúbizonyságom után az R program kódjaimat, egyedi funkcióimat értelmeztem, mutattam be röviden, amelynek megismertetése a későbbi számítások háttére miatt volt fontos. Az R alkalmazása Magyarországon még nem terjedt el számos kimagasló előnye ellenére, nem úgy, mint Európa többi részén.

A gyakorlati részben a kismintás (60 fős) kutatásom megalapozta a további kutatási irányok hátterét, amit a reprezentatív mintákon tudtam bizonyítani. Mind a három felhasznált adatbázis esetén a tényleges modellépítést alapvető statisztikai eszközök alkalmazása a minta megismerése és a sejtés megalapozása érdekében. A TÁRKI kérdőíven, amelynek elkészítésében én is tevékenyen részt vettem, alapuló reprezentatív minta segítségével egy bináris logisztikus regresszió alapuló modellt építettem (Logit1), amelynek négy szignifikáns magyarázó változója befolyásolja az eredményváltozó (Váltana-e szolgáltatót a fogyasztó?) feltételes valószínűségét. A kovariánsokat az áttekinthetőség és a következtetések levonása érdekében grafikonokon ábrázoltam. A jósági mutatók esetén a ROC görbe és a görbe alatti terület arra utalt, hogy a modell nem minden kimenetet képes jól megközelíteni, vagyis a modell fejlesztésére volt szükség. A megírt funkciókon alapuló algoritmus megalkotásával kerestem egy olyan modellt, amivel képes vagyok jól közelíteni az Y alakulását, úgy, hogy a modell jósága is kielégítő. Továbbá az egyes változók hozzájárulását is vizsgáltam. Így épült fel a szintén TÁRKI-s adatbázison alapuló Logit2 modell, amelynek magyarázó változói között szerepel a megye és az anyagi helyzet is, a mobil és a gázszolgáltató váltással kapcsolatos kérdések mellett. A Logit2 modell esetén egyéb rögzített

már tárgyalt paraméterek mellett a budapesti lakosok váltásának feltételes valószínűsége alacsonyabb, mint a Borsod megyei lakosoké, ám a váltás valószínűsége minden vizsgált megyében nő, ahogy az anyagi körülmények romlanak. A jósági mutatóknak megfelelt a modell. A más szolgáltatás váltással kapcsolatos kérdésre igen választ adó fogyasztók nagyobb eséllyel váltanak áram szolgáltatót is, amit a komplex szolgáltatást nyújtó vállalatoknak érdemes figyelembe venniük a fogyasztó megszerzése érdekében. Azaz a hipotézisemet, miszerint lakossági fogyasztók magatartását demográfiai jellemzők is befolyásolják nem tudtam elvetni.

A következő egységben a MEKH adatbázisra épülő mintát vizsgáltam, ahol az eredményváltozó szintén a szolgáltató változót jelenítette meg. A Logit3 modellben a magyarázó változók között szerepel az fogyasztó megítélése arról, hogy mennyire ügyfélbarát a szolgáltató, az áramkimaradással kapcsolatos kommunikáció fontossága, hogy melyik egyetemes szolgáltatóhoz tartozik a fogyasztó, valamint a neme. A jósági mutatók alapján a modell gyengébb, mint a Logit2 modell, viszont a területbeli és nembeli különbségek figyelemre méltóak. Azok közül a háztartási fogyasztók közül, akik nem tartják fontosnak az áramkimaradásokkal kapcsolatos előzetes kommunikációt és az egyetemes szolgáltatójuk ügyfélbarát hozzáállását rossznak értékelték, a férfiak nagyobb eséllyel váltának áramszolgáltatót, mint a nők. A váltási hajlandóság eszerint a modell szerint a kiemelt esetekben pedig az TITÁSZ területén a legmagasabb. A dolgozat utolsó részében röviden összefoglaltam a főbb következtetéseket.

Jövőbeni célom a fogyasztói tudatosság és fogyasztói elégedettség vizsgálata, kiterjesztése egy új kutatásban, amelyhez ez a dolgozat jó alapként szolgál. Összességében elmondható, hogy a hipotézisemet elfogadom, ugyanis a demográfiai változók hatással vannak a lakossági fogyasztók magatartására. Továbbá úgy gondolom, hogy a vállalatoknak érdemes megfontolni, hogy mennyiben koncentrálnak vizsgált tényezőkre.

IRODALOMJEGYZÉK

Szakirodalom

Adler, J. (2010) R in Nutshell, O'Reilly Media Inc.

András Zoltán (2013) Omnibusz 2013/06 adatok

Barna, J. (s.a.) Bevezetés a kvantitatív adatelemzésbe Elérhetőség: <http://barna.tatk.elte.hu/SPSS/orai%20anyagok/kereszttabla-elemzes.pdf> (Letöltve: 2013. november)

Ben-Akiva, M. (1999) Discrete Short Methods and Their Applications to Short Term Travel Decisions, Transportation Science Handbook, 1999 Elérhetőség: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.8438&rep=rep1&type=pdf> (Letöltve: 2013. november)

Cox, D.R. (1964) Analysis of transformation Elérhetőség: <http://fisher.osu.edu/~schroeder.9/AMIS900/Box1964.pdf> (Letöltve: 2013. november)

Cramer, J.S. (2002) The Origin of Logistic Regression, Faculty of Econometrics, University of Amsterdam, Discussion Paper, 2002 Elérhetőség: <http://dare.uva.nl/document/204> (Letöltve: 2013. november)

Fawcett, T. (2005) An introduction to ROC Analysis, Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA Elérhetőség: <http://people.inf.elte.hu/kiss/12dwhdm/roc.pdf> (Letöltve: 2013. november)

Fawcett, T., 2001. Using rule sets to maximize ROC performance. In: Proc. IEEE Internat. Conf. on Data Mining

Fiel, A., Miles, J. (2010) Discovering Statistics using with R, Sage Publication Inc, 2010

Flach, P., Wu, S., 2003. Repairing concavities in ROC curves. In: Proc. 2003 UK Workshop on Computational Intelligence. University of Bristol

Fülöp, P., (2002) A bináris logit modellek használatának és tesztelésének eszközei, Statisztikai Szemle, 2002. III. évfolyam Elérhetőség: http://www.ksh.hu/statszemle_archive/2002/2002_03/2002_03_261.pdf (Letöltve: 2013. november)

Gray R, Kovács, E. (2001) Az általánosított lineáris modell és biztosítási alkalmazásai, Statisztikai Szemle, 2001

Jászay Tamás (2010) Cégiskola című előadása alapján saját szerkesztés, ELMU– Nyrt.

Hajdu, O. (2003) Többváltozós statisztikai számítások, Budapesti Corvinus Egyetem

Hetesi, E., Rekettye, G. (2001) A lakossági ügyfelek lojalitásának mérése és elemzése egy magyar energiaszolgáltatónál, Vezetéstudomány, 32. évf. 9. sz., pp. 17-25. 1

- Hetesi E. (2002): A lojalitás mérési problémái a közüzemi szolgáltatóknál JATEPress 2002, Szeged, pp. 205-218.
- Hetesi E. (2002): A lojalitás mérési problémái a közüzemi szolgáltatóknál In: Hetesi E. (szerk.) 2002: *A közszolgáltatások marketingje és menedzsmentje* SZTE Gazdaságtudományi Kar Közleményei 2002. JATEPress 2002, Szeged, pp. 205-218
- Hosmer, D., Lemeshow, S. (1989): Applied logistic regression. John Wiley and Sons, New York
- Hunyadi, L. (2002) Grafikus ábrázolás a statisztikában, Statisztikai Szemle Elérhetőség: http://www.ksh.hu/statszemle_archive/2002/2002_01/2002_01_022.pdf (Letöltve: 2013. november)
- Hunyadi, L. (2004) A logisztikus függvény és a logisztikus eloszlás, Statisztikai Szemle 2004. X.-XI Elérhetőség: http://www.ksh.hu/statszemle_archive/2004/2004_10-11/2004_10-11_991.pdf (Letöltve: 2013. november)
- Hunyadi, L., Vita, L. (2010) Statisztika 1., Aula Tankönyvkiadó
- Hunyadi, L., Vita, L. (2012a) Következtető statisztika előadások, Hipotézis vizsgálat haladóknak, XX Egyetem
- Hunyadi, L., Vita, L. (2012b) A következtető statisztika alapjai közgazdászoknak
- Király, Z. (2012) Statisztika II. Asszociációs mutató számok, R-ben Elérhetőség: http://psycho.unideb.hu/munkatarsak/hidegkuti_istvan/targyak/Kiraly_Zoltan_Statisztika_2_jegyzet_1.pdf (Letöltve: 2013. november)
- Klemperer, P. (1987a) The competitiveness of Marktes with Switching costs, The RAND Journal of Economics, Vol, 18., No. 1.
- Klemperer, P. (1987b) Entry Deterrence in Marktes with Consumer Scwitching Cost, The Economica Journal, Vol 97
- Klemperer, P. (1989) Price Wars Caused by Switching Cost, The Review of Economis Studies, Vol. 56, No. 3.
- Kovács, E. (2006) Pénzügyi adatok statisztikai elemzése, Tanszék KFT. Budapest, 2006
- Long, S. (1997) Regression models for categorical and limited dependent variables, SAGE Publications, New York
- Maddala, G.S. (1995): Specification tests in limited dependent variable models. In: Maddala, G. S., Philips (szerk.), Advances in econometrics and quantitative economics: essays in honor of C.R. Rao. Basil Blackwell, Oxford.
- Pap Zoltán (2013) Adatok a villamosenergia-szolgáltatásról alkotott fogyasztói véleményekkel kapcsolatos felmérések - lakossági rész, Magyar Energetikai és Közmű Szabályozási Hivatal, Adatok 2009

- Pintér J., Rekettye G., (2008) A villamosenergia-szolgáltatással való elégedettség mérésekor alkalmazható kvantitatív módszerek és modellek, Szigma, XXXIX: (2008) 1.-2., Matematikai Közgazdaságtan
- Regionális Energiagazdasági Kutatóközpont (2007) Kisfogyasztók a liberalizált villamosenergia-piacon, E.On megbízásával készült a Budapesti Corvinus Egyetem segítségével, Sugár András és Kaderják Péter vezetésével
- Rekettye, Orosdy, Tersztyánszky (1996-2004) A villamosenergia-szolgáltatásról alkotott fogyasztói vélemények (1996.-2004.) évi vizsgálatának eredményei, Energiagazdálkodás, PTE KTK Marketing Tanszék, Marketing oktatók III. éves konferenciája, Tatabánya 1997, Marketing és Menedzsment (1997)
- Shy, O (2001) The Economics of Network Industries, Cambridge University Press
- Sugár András (2011) A piacsabályozás elméleti és gyakorlati aspektusai a közszolgáltató szektorokban, elsősorban az energiaszektor árszabályozása példáján Ph. D értekezés Budapest, Budapesti Corvinus Egyetem, Statisztika Tanszék
- TÁRKI (2013a) Omnibusz módszertani útmutató, TÁRKI
- TÁRKI (2013b) Röviden a TÁRKI Omnibusz vizsgálatról, Elérhetőség: http://www.tarki.hu/hu/services/su/omnibusz_ismerteto.pdf (Letöltve: 2013. június)
- Tétényi, L. (2013) Mikroökonometria R-ben előadások alapján, XX Egyetem
- Veall, Zimmermann: Pseudo-R² Measures for Some Common Limited Dependent Variable Models Sonderforschungsbereich 386, Paper 18 (1996) Elérhetőség: http://epub.ub.uni-muenchen.de/1421/1/paper_18.pdf (Letöltve: 2013. november)
- Szerző (2012) Az egyetemes szolgáltatás helyzetének vizsgálata a magyar villamosenergia-szektorban a szolgáltatóváltások és az új trendek tükrében
- (S.n.) (2010) Goodness of fit in Logistic Regression, Medical Elérhetőség: <http://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logfit.pdf> (Letöltve: 2014. január)
- (S.n.) (2012a) Analytics Trainings Elérhetőség: <http://analyticstrainings.com/?p=101> (Letöltve: 2014. január)
- (S.n.) (2013a) R programnyelv, objektum orientált programozási nyelvek, ELTE Informatikai Kar, anyagok Elérhetőség: <http://nyelvek.inf.elte.hu/leirasok/R/index.php?chapter=10> (Letöltve: 2013. október)

Jogsabályok

1994. évi XLVIII. törvény a villamos energia termelésről, szállításról és szolgáltatásról, <http://www.complex.hu/kzldat/t9400048.htm/t9400048.htm>

2001. évi CX. törvény a villamos energiáról, egységes szerkezetben a végrehajtásról szóló 180/2002. (VIII.23.) Korm. Rendelettel,
http://www.gvh.hu/domain2/files/modules/module25/pdf/2001_evi_CX_torveny.pdf
2007. évi LXXXVI. törvény a villamos energiáról,
http://www.complex.hu/jr/gen/hjegy_doc.cgi?docid=A0700086.TV#lbj8param,
- 2003/54/EK irányelv a villamos energia belső piaci szabályozásáról és a 96/92/EK hatályon kívül helyezéséről, <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=DD:12:02:32003L0054:HU:PDF>
- 2009/72/EK irányelv a villamos energia belső piaci szabályozásáról és a 2003/54/EK hatályon kívül helyezéséről, <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:211:0055:0093:HU:PDF>
- 273/2007 Kormány rendelet a villamos energiáról szóló 2007. évi LXXXVI. törvény egyes rendelkezéseinek végrehajtásáról,
http://www.masz.co.hu/pdf/273.%282007%29._%28X._19.%29_Korm._rendelet.pdf
- 4./2011. (I.31.) NFM rendelet a villamos energia egyetemes szolgáltatás árképzéséről,
http://www.complex.hu/jr/gen/hjegy_doc.cgi?docid=A1100004.NFM