REPUBLIKA E SHQIPËRISË
UNIVERSITETI POLITEKNIK I TIRANËS
FAKULTETI INXHINIERISË MATEMATIKE DHE INXHINIERISË FIZIKE
DEPARTAMENTI INXHINIERISË MATEMATIKE

DISERTACION

Për marrjen e gradës shkencore

"DOKTOR"

Paraqitur nga

MSc. Erjola KASTRATI (CENAJ)

MODELIMI MATEMATIK I TË DHËNAVE KATEGORIKE DHE ZBATIME

Udhëheqës Shkencor

Prof. Asoc. Dr. Luela PRIFTI

Tiranë, 2017

REPUBLIKA E SHQIPËRISË
UNIVERSITETI POLITEKNIK I TIRANËS
FAKULTETI INXHINIERISË MATEMATIKE DHE INXHINIERISË FIZIKE
DEPARTAMENTI INXHINIERISË MATEMATIKE

DISERTACION

Për marrjen e gradës shkencore

"DOKTOR"

Paraqitur nga

MSc. Erjola KASTRATI (CENAJ)

MODELIMI MATEMATIK I TË DHËNAVE KATEGORIKE DHE ZBATIME

Udhëheqës Shkencor: Prof. Asoc. Dr. Luela PRIFTI

Mbrohet më datë ___/___/ 2017 para jurisë:

1. Prof. Asoc. Dr. Ligor NIKOLLA	Kryetar
2. Prof. Dr. Shpëtim LEKA	Anëtar, Oponent
3. Prof. Asoc. Dr. Shkëlqim KUKA	Anëtar, Oponent
4. Prof. Dr. Viktor KABILI	Anëtar
5. Prof. Dr. Fatmir HOXHA	Anëtar

FALENDERIME!

Përfundimit të kësaj pune disa vjeçare do ti shtonte vlera shprehja e mirënjohjes dhe përcjellja e disa falenderimeve për të gjithë ata që përkrahën dhe motivuan akoma më tej dëshirën dhe përpjekjen time për këtë kualifikim akademik.

E falenderoj me shumë mirënjohje udhëheqësen time shkencore Prof. Asoc. Dr. Luela Prifti për përkushtimin dhe mbështetjen që më ka dhënë gjatë gjithë kohës së pregatitjes së këtij punimi.

Falenderoj profesorin e nderuar Prof. Dr Shpetim Shehu për këshillat, sygjerimet e inkurajimin për të përfunduar në kohë të gjithë proçesin e doktoraturës.

Falenderoj kolegët e mi të departamentit të Inxhinierisë Matematike për mbështetjen dhe këshillat e tyre.

Përcjell shumë falenderime dhe do i jem gjithnjë mirënjohse familjes time në mënyrë të veçantë prindërve të mi, për ndimën dhe mbështëtjen e pakursyer e të palodhur gjatë formimit tim profesional akademik.

Falenderoj përzemërsisht bashkëshortin tim Avenir dhe të shtrejtin tim Sion, për dashurinë, durimin, mbështetjen dhe qetësinë që më kanë dhuruar.

PËRMBAJTJA

LISTA E FIGURAVE……...v

LISTA E TABELAVE……...……...vi

PËRMBLEDHJE……..……1

ABSTRACT……...…….2

KAPITULLI 1……3

TË DHËNAT KATEGORIKE DHE SHPËRNDARJET PROBABILITARE TË TYRE

……...…...3

1.1 Hyrje……..3

1.2 Të dhënat kategorike…..……3

1.3 Ndryshoret e rastit …….5

1.3.1 Ndryshoret e rastit diskrete dhe të vazhduara ……...5

1.4 Shpërndarjet probabilitare për të dhënat kategorikes……5

1.4.1 Shpërndarja Binomiale……..……..6

1.4.2 Shpërndarja Multinomiale……..6

1.4.3 Shpërndarja e Puasonit…….....7

1.4.4 Mbidispersioni (Overdispersion)……....8

1.4.5 Lidhja midis shpërndarjes së Puasonit dhe Multinomiale……..……9

1.5 Vlerësimet Satistikore për të dhënat kategorike…….…….10

1.5.1 Funksioni i përgjasisë dhe vlerësimet me përgjasinë maksimale……...……..10

1.5.2 Veti të Vlerësuesit të përgjasisë maksimal (VPM)……...……11

1.5.3 Funksioni i përgjasisë për parametrin binomial……...……12

1.5.4 Testi për një parametër binomial……..13

1.5.5 Intervalet e besimit për një parametër …….……14

1.6 Statistika inferenciale për parametrat multinomial……..16

1.6.1 Vlerësimet për parametrat multinomial……16

1.6.2 Testi për parametrin Multinomial……17

1.7 Metadat statistikore Bayes-iane për të dhënat kategoreike dhe teorema e Bayes……18

1.7.1 Shpërndarja prior për një parametër Binomial……..19

1.7.2 Vlerësimi Bayesian për parametrin binomial……...21

1.7.3 Vlerësimi i parametrave multinomial……...22

1.8 Përfundime……...24

KAPITULLI 2……..25

TABELAT E KONTIGJENCËS DHE STRUKTURA PROBABILITARE E TYRE…….25

2.1 Hyrje …….....25

2.2 Struktura probabilitare për tabelat e kontigjencës……...25

2.2.1 Tabelat e kontigjencës dhe shpërndarjet e tyre …….....25

2.2.2 Tabelat e kontigjencës 2×2……...26

2.2.3 Tabela e kontigjencës m×n.....29

2.2.4 Pavarësia statistikore e tabelave të kontigjencës 2x2.....30

2.2.5 Pavarësia statistikore e tabelave të kontigjencës 2x3……...31

2.3 Krahasimi i dy raporteve ……33

2.3.1 Diferenca e raporteve……33

2.3.2 Raporti i mundësive (shanseve)…….....33

2.3.3 Lidhja midis Raportit të mundësive dhe Riskut Relativ.....34

2.3.4 Pavarësia statistikore e tabelave të kontigjencës mxn.....34

2.4 Matricat e kontigjencës.....36

2.4.1 Pavarësia e tabelës së kontigjencës 2x2.....36

2.4.2 Pavarësia e tabelës së kontigjencës 3x3.....37

2.4.3 Shpërndarjet margjinale dhe matrica e vlerave të pritshme.....38

2.5 Shpërbërja e matricave të mbetjes.....40

2.5.1 Shpërbërja e matricave 2×2.....41

2.6 Shpërbërja e matricave 3×3.....43

2.6.1 Matrica 3 × 3 me rang 2.....43

2.6.2 Matrica 3 × 3 me rang 3.....45

2.7 Tabelat e kontigjencës me tre hyrje.....46

2.8 Vlerësimi i qelizave probabilitare në tabelat e kontigjencës.....48

2.8.1 Vlerësimi i parametrave binomial.....48

2.8.2 Vlerësimi i qelizave probabilitare multinomiale.....49

2.8.3 Vlerësimi i parametrave të modelit loglinear në tabelat me dy hyrje.....50

2.9 Përfundime.....52

KAPITULLI 3……..53

MODELI I REGRESIT LOGJISTIK TË THJESHTE DHE TË SHUMËFISHTË…….53

3.1 Hyrje…….....53

3.2 Modeli i Regresit të thjeshtë logjistik …….53

3.2.1 Përshtatja e Modelit të Regresit Logjistik…….....59

3.2.2 Kontrolli i rëndësisë së koeficientëve……...61

3.2.3 Vlerësimi i intervalit të besimit…….....64

3.3 Modeli i Regresit të shumëfishtë logjistik……...65

3.3.1 Përshtatja e Modelit të Regresit Logjistik të Shumëfishtë……66

3.3.2 Kontrolli i Rëndësisë së Modelit……..67

3.4 Përfundime……..68

KAPITULLI 4……..70

MODELIMI I TË DHËNAVE ME REGRESIN LOGJISTIK TË SHUMËFISHTË……70

4.1 Hyrje……70

4.2 Historik i shkurtër mbi zhvillimin e sigurimeve shoqërore në Shqipëri……..70

4.3 Të dhënat……..74

4.4 Modelimi i të dhënave me regresin e shumëfishtë logjistik……75

4.4.1 Rasti i parë (Varësia midis dy variablave kategorik)……76

4.4.2 Rasti i dytë (Varësia e ndryshores kategorike me një ndryshore të vazhdueshme)……...78

4.4.3 Rasti i tretë (Varësia e ndryshores kategorike me ndryshore të vazhdueshme, kategorike)……79

4.5 Përfundime …….....85

REFERENCAT……86

LISTA E FIGURAVE

Figura 1.1: Funksioni i përgjasisë dhe logaritmi i funksionit të përgjasisë……15

Figura 3.1: Prania dhe mungesa e sëmundjes për 60 individë……...55

Figura 3.2: Raporti i pranisë së sëmundjes në varësi të moshës ……...56

Figura 3.3: Grafiku i funsionit logjistik…….57

Figura 4.1: Kontribuesit në skemën e sigurimeve shoqërore sipas sektorve…….72

Figura 4.2: Struktura e financimit të ISSH……72

Figura 4.3: Shpenzimet në vitin 2016……73

Figura 4.4: Numri i kontribuesve urbanë e rurale në vite……..74

LISTA E TABELAVE

Tabela 2.1: Tabela e kontigjencës për dy karakteret A dhe B……...25

Tabela 2.2 Tabela e kontigjencës me përmasë 2×2…….....26

Tabela 2.3 Tabela e kontigjencës me përmasë mxn……..29

Tabela 2.4 Tabela e kontigjencës me përmasë 2×3.....31

Tabela 3.1 Mosha dhe gjinia prania dhe mungesa e sëmundjes tek 85 persona……54

Tabela 3.2: Denduritë e sëmundjes sipas moshës……..56

Tabela 3.3: Përshkrimi i modelit të regresit logjistik dhe regresit linear……..58

Tabela 3.4: Vlerat e Modelit të Regresit Logjistik kur variabli i pavarur është kategorike……..60

Tabela 3.5: Përshkrimi i tiparave për bashkësinë e të dhënave…….61

Tabela 4.1: Numri i kontribuesve në skemën e Sigurimeve Shoqërore sipas sektorit të punësimit……71

Tabela 4.2: Struktura e shpenzimeve në vitin 2016……...72

Tabela 4.3: Kontribues të sigurimeve shoqërore në vite……73

Tabela 4.4: Shpërndarja e individëve sipas pjesmarrjes në skemën e sigurimeve shoqërore……74

Tabela 4.5: Përshkrimi i tipareve për bashkësinë e të dhënave……75

Tabela 4.6: Tabela e kontigjencës për shpërndarjen e individëve të përfshirë në skemë sipas gjinisë……76

Tabela 4.7: Testi Omnibus për koeficientët e modelit (rasti i parë)……..77

Tabela 4.8: Treguesit e modelit te regresit logjistik (rasti i parë)……..78

Tabela 4.9: Tabela e koeficientëve të ndryshoreve në ekuacion (rasti i parë)……...78

Tabela 4.10: Testi Omnibus për koeficientët e modelit (rasti i dytë)……79

Tabela 4.11: Treguesit e modelit te regresit logjistik (rasti i dytë)……79

Tabela 4.12: Tabela e koeficientëve të ndryshoreve në ekuacion (rasti i dytë)…….79

Tabela 4.13: Kodimi i ndryshoreve kategorike……..80

Tabela 4.14: Testi Omnibus për koeficientët e modelit (rasti i tretë)……80

Tabela 4.15: Treguesit e modelit të regresit logjistik (rasti i tretë)……81

Tabela 4.16: Tabela e klasifikimit…….81

Tabela 4.17: Tabela e koeficientëve të ndryshoreve në ekuacion (rasti i tretë)……..82

PËRMBLEDHJE

Në ditët tona përdorimi i ndryshoreve kategorike haset në shumë analiza statistikore, sidomos në ato që lidhen me zbatime në shkencat sociale, në mjekësi, në fushën bankare, etj. Analiza e të dhënave kategorike ka nisur në 1960. Përmirësimi i metodave bazë dhe ndërtimi i metodave të reja vazhdon dhe në ditët tona.

Qëllimi i këtij studimit është modelimi i të dhënave kategorike bazuar në metoda statistikore dhe zbatimi i tyre me të dhëna reale në SPSS.

Studimi është i organizuar në katër kapituj. Në kapitullin e parë paraqiten ndryshoret kategorike dhe shpërndarjet probabilitare më të rëndesishme që lidhen me to. Në këtë kapitull gjithashtu paraqitet lidhja midis shpërndarjeve probabilitare, vlerësimi i parametrave të tyre me metodën e përgjasisë maksimale, vlerësimi intervalor si dhe kontrolli i hipotezave. Paraqiten vlerësimet Bayes-iane për të dhënat kategorike, për parametrat Binomial dhe Multinomial.

Në kapitullin e dytë trajtohen tabelat dhe matricat e kontigjencës me dy-hyrje. Tabelat e kontigjencës me 2-hyrje mund të zgjerohen në tabela të kontigjencës me shumë hyrje për ndryshoret multinomiale. Ndërtohet matrica e kontigjencës, elementët e së cilës janë të barabarta me vlerat korresponduese të tabelës së kontigjencës duke përjashtuar vlerat margjinale. Tregohet se kur një tabelë kontigjence merret si matricë atëherë për pavarsinë statistikore midis ndryshoreve rol kryesor luan rangu i matricës. Në pjesën e fundit të këtij kapitulli trajtohen vlerësimet Bayes-iane për vlerësimin e qelizave probabilitare në tabelat e kontigjencës.

Në kapitullin e tretë trajtohen modelet e regresit logjistik të thjeshtë, regresit logjistik të shumëfishtë, si pjesë përbërëse e analizës së të dhënave. Realizohet përshtatja e modelit, kontrolli i rëndësisë së koeficientëve dhe vlerësimi intervalor i tyre. Për problemin e parashikimit të pranisë apo mungesës së sëmundjes tek individët analizohet dhe intepretohet modeli i regresit logjistik të thjeshtë që ndërtohet nga përdorimi i SPSS.

Në kapitullin e katërt është realizuar modelimi i të dhënave reale me regresin logjistik të shumëfishtë duke përdorur programin SPSS. Baza e të dhënave përmban 3310 të dhëna të marra nga The World Bank. Modeli i ndërtuar analizon parashikimin e probabilitetit të mos përfshirjes së një individi të punësuar në Skemën e Sigurimeve Shoqërore.

Përfundimet janë vendosur në pjesën e fundit të çdo kapitulli.

ABSTRACT

Nowadays categorical variables are used extensively in statistical analysis, mostly related to social sciences, medicine, banking field etc. The analyse of the categorical variables has started in 1960. The improvement of the basic methods and the defining of the new ones continues still in our days.

This study aims to model categorical data based on statistical methods and to realize applications with real data using SPSS. The study is organized in four chapters.

The first chapter present the categorical variables and their probability distributions. Also in this chapter is presented the connection between probability distributions, the maximum likelihood method, interval estimation and statistical hypothesis testing. In the last part is presented the Bayes estimators of the categorical data and of the binomial and multinomial parameters.

In the second chapter are treated the contingency tables and matrixs two-way. Two-way contingency tables can be expanded in multi way contingency tables for multinomial variables. A contingency matrix is determined by the elements which are equal to the corresponding elements of the contingency tables values excluding marginal values. When a contingency table is considered as a matrix then the matrix rank plays the main role for the independence of the statistical variables.

The third chapter deals with the simple logistic regression methods, multiple logistic regressions as an important part of data analysis. The simple logistic regression model is being analyzed for the problem of predicting presence or absence of disease in individuals. The results are derived from the use of the SPSS.

In the fourth chapter is realized the modeling of the real data with the multiple logistic regression using the program SPSS. The database contains 3310 data received from The World Bank. The defined model analyzes the probability prediction of not including an employed person in the social insurance scheme. The conclusions are presented in the last part of each chapter.

Keywords: Categorical data, Contingency table, Logistic Regression

Kapitulli 1

TË DHËNAT KATEGORIKE DHE SHPËRNDARJET PROBABILITARE TË TYRE

1.1 Hyrje

Në këtë kapitull trajtohen ndryshoret kategorike, shpërndarjet probabilitare më të rëndësishme që lidhen me to, si dhe paraqitet lidhja midis shpërndarjeve probabilitare. Është trajtuar dhe vlerësimi i parametrave të tyre me metodën e përgjasisë maksimale, futet kuptimi i funksionit të përgjasisë që përmban parametrat i cili quhet *kernel*, vlerësimi intervalor dhe kontrolli i hipotezave sipas Agrestit (2002). Në shembujt e paraqitur realizohen zbatime të metodave të përmendura. Jepen vlerësimet Bayes-iane për të dhënat kategorike, dhe tregohet se metodat Bayes-iane janë integrim i teoremës së Bayes-it. Tregohet se peshat relative të mundësisë dhe priori përcaktohen nga varianca e shpërndarjes sipas Agrestit et al. (2005). Paraqiten gjithashtu dhe vlerësimet Bayes-iane për parametrin binomial dhe multinomial.

1.2 Të dhënat kategorike

Përkufizim 1.2.1: Një ndryshore kategorike përmban një bashkësi kategorish sipas një shkallë matjeje.

Ajo lejon renditjen e karakteristikave të ndryshores në kategori. Për shëmbull: Filozofia politike është disa llojesh: liberale, e moderuar ose konservative. Diagnoza që lidhet me kancerin e gjirit bazohet në përdorimin e mamogramës për këto kategori: normale, e parrezikshme, e mundshme, dyshuese dhe kritike etj.

Zhvillimi i metodave për ndryshoret kategorike u nxit nga studimet kërkimore në shkencat sociale dhe mjekësore. Sot ato përdoren gjerësisht në shkencat sociale për vlerësimin e qëndrimeve dhe opinioneve, në shkencat mjekësore përdoren për vlerësimin e rezultateve, në qoftë se një trajtim mjekësor është i suksesshëm, në analizen e të sjellurit (per shembull: llojet sëmundjeve mendore, me këto kategori si: skizofrenia, depresioni, nervozizimi), në epidemiologji dhe në shëndet publik (per shembull: metodat kontraceptive në një marrëdhënie: kondom, pilulë, IUD etj), në zoologji (per shembull: aligatorët si parapëlqim për ushqim primar,

të kategorisë së peshqëve, jovertebrorët, zvarranikët), në arsim (per shembull: përgjigjet e studentevë të pyetjeve të provimit, të kategorisë e sakte dhe gabim), në marketing (per shembull: preferencat e konsumatorëve për markën kryesore të një produkti, i kategorisë A, B dhe C).

Ndryshoret kategorike klasifikohen si më poshtë:

Përkufizim 1.2.2: Ndryshoret që marrin vlera jo sipas një renditje të zakonshme janë quajtur nominale.

Këto ndryshore janë emra, etiketime, kategori, gjini etj. Shembuj janë: besimi fetar (i kategorisë ortodoks, katolik, protestant, çifut, mysliman etj), llojet e transportit publik (automobilë, biçikletë, autobus, metro etj), lloji i muzikës (klasike, kantri, folk, xhaz, rok etj), dhe vendbanimi (apartament, godinë në bashkëprononësi, shtëpi private, etj). Për ndryshoret nominale analiza statistikore nuk varet nga renditja. Mbi to nuk mund të kryhen veprime aritmetike. Për shembull gjiinia: mashkull, femer; nuk ka renditje dhe as veprime aritmetike.

Përkufizim 1.2.3: Ndryshoret që marrin vlera sipas kategorive të renditura quhen ndryshore të renditura.

Ato mund të renditen sipas një kriteri. Edhe me këto të dhëna nuk mund te kryhen veprime aritmetike. Shembuj janë kursi shkollor (viti I, II, III, ….), klasa shoqërore (e lartë, e mesme, e ulët), filozofia politike (liberale, e moderuar, konservative) dhe gjendja shpirtërore (e mirë, serioze, e rrezikshme etj). Ndryshoret e renditura kanë kategori të renditura. Për shembull një individ i klasifikuar si i moderuar i cili është më shumë liberal se sa një individ i klasifikuar si konservator, asnjë vlerë numerike nuk përcakton se sa liberal është një individ. Metodat për ndryshoret e renditura përdorin renditjen e kategorive.

Perkufizim 1.2.4: Ndryshore intervalore është ajo ndryshore vlerat e së cilës ndodhen brenda një intervali (të përcaktuara midis dy vlerave numerike).

Është e ngjashme me ndryshoren e renditur. Për shembull, niveli i presionit të gjakut, jetëgjatësia funksionale e televizorëve, kohëzgjatja e periudhës në burg, të ardhurat vjetore etj. Mënyra se si një ndryshore është përcaktuar lidhet me klasifikimin e saj. Për shembull: “shkollimi” është ndryshore nominale kur njehsohet si shkollë publike ose shkollë private; ai është ndryshore e renditur kur përdoren kategoritë: nëntë vjeçare, shkollë e mesme, bachelor, master dhe doktoraturë; ajo është një ndryshore me vlera në interval kur matet nga numri i viteve të shkollimit, si numra të plotë 0,1,2, ...

Metodat statistikore për ndryshoret nominale mund të përdoren për ndryshoret e renditura ndërsa ato për ndryshoret e renditura nuk mund të përdoren për ndryshoret nominale. Është më mirë të përdoren metodat e duhura në shkallën reale. Metodat e përdorura për ndryshoret intervalore kanë një numër të vogël të vlerave të mundshme (si psh, shkollimi i matur < 10 vjet, 10 – 12 vjet, > 12 vjet).

1.3 Ndryshoret e rastit

Përkufizim 1.3.1: Ndryshore rasti X quhet çdo funksion real X i përcaktuar në hapësirën Ω pra

X: Ω → R.

Ndryshoret e rastit i shënojmë me shkronjat X, Y, Z, … etj. Gjatë studimit të një ndryshoreje të rastit problem më i rëndësishëm është gjetja e probabiliteteve të vlerave të ndryshme të saj. Problem i cili zgjidhet sepse vlerat e ndryshores së rastit përcaktohen nga rezultatet e provës, ku kemi të përcaktuar një probabilitet P. Shënojmë $\mathcal{H} = \{x_1, x_2, \dots, x_n\}$ bashkësinë e vlerave të mundshme të n.r **X**.

1.3.1 Ndryshoret e rastit diskrete dhe të vazhduara

Përkufizim 1.3.2: Ndryshorja e rastit X quhet diskrete në qoftë se bashkësia $\mathcal{H}$ është të shumtën e numërueshme.

Shembuj janë: numri i fëmijëve në familje, numri i studentëve në një klasë, numri i herëve që shkojmë në plazh etj.

Përkufizim 1.3.3: Ndryshorja e rastit X quhet e vazhduar në qoftë se gjendet një funksion$f(x) \geq 0$ (për cdo $x \in R$) që për cdo bashkësi $B \subset R$ të kemi:

$$P(X \in B) = \int f(x)dx$$

ku f(x) quhet densitet i probabiliteteve të n.r. X.

Përkufizim 1.3.4: Ndryshoret e rastit X dhe Y quhen të pavarura në qoftë se për $\forall\, x, y \in R$

$$P(X < x; Y < y) = P(X < x) \cdot P(Y < y)$$

1.4 Shpërndarjet probabilitare për të dhënat kategorike

Analiza e të dhënave kërkon supozime rreth mekanizmave të rastit që gjenerojnë të dhënat. Për modelet e regresit me përgjigje të vazhdueshme, shpërndarja normale luan një rol të rëndësishëm. Në këtë paragraf studiohen tre shpërndarje kryesore për të dhënat kategorike:

a) shpërndarja Binomiale, b) shpërndarja Multinomiale, c) shpërndarja e Puasonit

1.4.1 Shpërndarja Binomiale

Shumë aplikime i referohen një numri të caktuar *n* të vrojtimeve. Kryhet prova e cila ka vetëm dy rezultate "sukses" dhe "mossukses". Shënojme X ndryshoren e rastit me bashkësi vlerash {0,1} ku me "1" kemi shënuar suksesin dhe me "0" mossuksesin. Ndryshorja e rastit X ka shpërndarje Bernuli në qoftë se:

$$P(X=x)=\begin{cases} p \text{ në } qoftë \text{ se } x=1 \\ 1-p \text{ në } qoftë \text{ se } x=0 \end{cases} \qquad (1.1)$$

Pritja matematike $E(x)=1\times p+0\times(1-p)=p$, dhe varianca $Var(x)=p\cdot(1-p)=p\cdot q$

Supozojmë se përsërisim *n*-herë provat e pavarura të Bernulit. Shënojmë me **X** n.r. që paraqet numrin e "sukseseve" me këto n-prova. Kuptohet se $\mathbf{X}=\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_\mathrm{n}$ ku X_i ($\forall \mathrm{i}=1,2,\dots,\mathrm{n}$) janë n.r. të pavarura të Bernulit. Cdo rezultat i provës ka x- suksese dhe $(n-x)$-mos suksese. Numri i sukseseve $=\sum_{i=1}^{n}X_i$, ka shpërndarje Binomiale me parametër n dhe p, e përcaktur nga X$\sim B(n,p)$.

$P(X=x)=C_n^x\cdot p^x\cdot q^{n-x}$, $\forall x=0,1,2,\dots,n$, $E(X)=np, Var(X)=npq$

1.4.2 Shpërndarja Multinomiale

Disa prova kanë më shumë se dy rezultate të mundshme. Supozojmë që secila nga *n* provat e pavarura dhe identike mund të ketë rezultate në secilën nga *c* kategoritë. Le të jetë x_{ij}= 1 në qoftë se prova *i* ka rezultate në kategorinë *j* përndryshe x_{ij} = 0,

pra $x_{ij}=\begin{cases} 1 \text{ në } qoftë \text{ se } i=j \\ 0 \text{ në } qoftë \text{ se } i\neq j \end{cases}$

Kështu që, $x_i=(x_{i1},x_{i2},\dots,x_{ic})$ përfaqëson një provë multinomiale me $\sum_j x_{ij}=1$.

Për shembull, (0,0,1,0) paraqet rezultate në kategorinë 3 nga katër kategori të mundshme. Le të jetë x_{ic} linearisht i varur nga të tjerët dhe $n_j=\sum_i x_{ij}$ tregon numrin e provave që kanë rezultate në kategorinë *j*. Numërimet $(n_1,n_2,\dots,n_c)$ kanë *shpërndarje multinomiale.*

Le te jetë $p_j=P(X_{ij}=1)$ probabiliteti i rezultateve në kategorinë *j* për çdo provë.

Densiteti probabilitar është $p(n_1,n_2,\dots,n_{c-1})=\left(\frac{n!}{n_1!n_2!\dots n_c!}\right)p_1^{n_1}p_2^{n_2}\dots p_c^{n_c}$ (1.2)

Meqënëse $\sum_j n_j=n$, kjo është $(c-1)$ përmasore, me $n_c=n-(n_1+n_2+\cdots+n_{c-1})$. Shpërndarja binomiale është rast i veçantë me $c=2$. Për shpërndarjen multinomiale,

E(n_j) = n, Var(n_j) = np_j(1-p_j), cov(n_j, n_k) = -np_jp_k.

1.4.3 Shpërndarja e Puasonit

Ndonjëherë të dhënat numerike nuk merren nga një numër i fiksuar *i* provash. Për shembull, nëse shënoj me X = numrin e vdekjeve për shkak të aksidenteve automobilistike në autostradat në Itali gjatë javës së ardhshme, për *x* nuk ka kufi të sipërm *n*. Meqënëse *x* duhet të jetë një numër i plotë jonegative shpërndarja e tij probabilitare duhet të jetë në zonë. Shpërndarja më e thjeshtë për të është ajo e Puasonit. Probabilitetet e saj varen nga një parametër i vetëm, nga pritja matematike μ.

Le të jetë X ndryshore rasti diskrete me bashkësi vlerash {0,1,2,….}. Shpërndarja probabilitare është

$$P(x) = \frac{e^{-\mu}\mu^x}{x!}, x = 0, 1, 2, \ldots \text{ dhe } \mu > 0 \qquad (1.3)$$

Ajo plotëson kushtin që pritja matematike dhe varianca janë: $E(X) = var(X) = \mu$. Ajo ka modë të barabartë me pjesën e plotë të pritjes μ. Shpërndarja i afrohet asaj normales kur μ rritet. Shpërndarja e Puasonit përdoret për llogaritjen e ngjarjeve që ndodhin rastësisht në kohë apo hapësirë, kur rezultatet në periudha të veçanta janë të pavarura. Ajo gjithashtu merret si një përafrim i shpërndarjes binomiale kur *n* është e madhe dhe p është e vogël Leka (2004) këtë e tregon qartë dhe teorema e mëposhtme ,

Teorema 1.4.1: Në qoftë se X$\sim B(n,p)$, me supozimin se n-ja është e madhe dhe p-ja është shumë e vogël, duke shënuar μ = np, tregohet se $X \to P(\mu),\ n \to \infty$.

Vërtetë:

$$P(X=k) = C_n^k \cdot p^k \cdot q^{n-k} = \frac{D_n^k}{k!} \cdot \left(\frac{\mu}{n}\right)^k \cdot \left(1-\frac{\mu}{n}\right)^{n-k}$$

$$= \frac{\mu^k}{k!} \cdot \left(1-\frac{\mu}{n}\right)^{-k} \cdot \left(1-\frac{\mu}{n}\right)^n \cdot \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}$$

$$\cong e^{-\mu} \cdot \frac{\mu^k}{k!}$$

sepse $\left(1-\frac{\mu}{n}\right)^{-k} \cong 1; \left(1-\frac{\mu}{n}\right)^n \cong e^{-\mu}, \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k} \cong 1$

Pra $P(X=k) \cong e^{-\mu} \cdot \frac{\mu^k}{k!}$ domethënë $X \to P(\mu), n \to \infty$.

Një veti e shpërndarjes se Puasonit është se dispersioni i saj është i barabartë me pritjen matematike. Zgjedhja paraqet ndryshime kur mesatarja e tyre është e lartë.

Shembull 1.1.1

Në qoftë se secili nga 50 milion njerëz që do të ngasin makinën në Itali javën e ardhshme do të konsiderohet si një provë e pavarur me të njetin probabilitet 0.000002 për të vdekur në një aksident fatal atë javë atëherë numri i vdekjeve X ka shpërndarje B ~ (50 000 000, 0.000002), me parameter μ = np = 50 000 000×0.000002 = 100. Pra, numri mesatar i aksidenteve fatale javore është i barabartë me 100. Ndryshueshmëria më e madhe ndodh në llogaritje të përjavshme krahasuar me rastin kur mesatarja është e barabartë me 10.

1.4.4 Mbidispersioni (Overdispersion)

Në praktikë, të dhënat numerike të vrojtuara shpesh shfaqin ndryshueshmëri duke e kaluar atë që parashikohet nga shpërndarja binomiale te e Puasonit. Kjo dukuri njihet si *mbidispersion*.

Në paragrafin 1.4.2 supozohet se çdo person ka të njëjtin probabilitet të vdekjes në një aksident fatal në javën e ardhshme. Këto probabilitete ndryshojnë për shkak te faktorëve të tillë si: sasia e kohës së shpenzuar gjatë ngarjes së makinës, nëse personi vendos rripin e sigurimit dhe vendndodhja gjeografike. Ndyshime të tilla shkaktojnë fatkeqësi të cilat shfaqin më shumë ndryshime sesa parashikimet nga modeli i Puasonit.

Supozohet se X është një ndryshore rasti me variacion $var(X|\,\mu)$ për një μ të dhënë, por μ ndryshon për shkak të faktorëve të pamatur të tilla si ato që përmendëm më sipër.

Le të jetë $\theta = E(\mu)$, atëherë,

$$E(X) = E[E(X|\mu)], \quad var(X) = E[var(X|\,\mu)] + var[E(X|\,\mu)].$$

Kur X është i kushtëzuar nga shpërndarja e Puasonit (jepet μ), atëherë

$$E(X) = E(\mu) = \theta$$

dhe

$$var(X) = E(\mu) + var(\mu) = \theta + var(\mu) > \theta.$$

Supozimi i shpërndarjes së Puasonit për një ndryshore numerike (count variable) është shumë i thjeshtë për shkak të faktoreve që shkaktojnë mbidispersionin. Një shpërndarje binomiale negative është shpërndarje e lidhur me të dhëna numerike të cilat lejojnë dispersionin ta tejkaloj mesataren. Analiza e shpërndarjeve binomiale ose multinomiale janë ndonjëherë të pavlefshme

për shkak të mbidispersionit. Kjo ndodh sepse shpërndarja e vërtetë është një përzierje e shpërndarjes binomiale me një parametër që ndryshon për shkak të ndryshoreve të pamatura.

1.4.5 Lidhja midis shpërndarjes së Puasonit dhe Multinomiale

Me të dhënat e marra nga aksidentet autmobilistike, ajrore dhe hekurudhore gjatë një viti në Itali shënojmë:

X_1 = numri i personave të cilët kanë vdekur nga aksidentet automobilistike të Europës,

X_2 = numri i personave që kanë vdekur nga aksidentet ajrore,

X_3 = numri i personave që kanë vdekur nga aksidentet hekurudhore.

Modeli i Puasonit (X_1, X_2, X_3) i trajton këto ndryshore të rastit si ndryshore rasti Puasoni të pavarura me parametra respektive (μ_1, μ_2, μ_3). Densiteti probabilitar për $\{X_i\}$ është prodhimi i tre shpërndarjeve probabilitare të formës (1.3). Në përgjithësi $n = \sum X_i$ ka një shpërndarje të Puasonit, me parametër $\Sigma\ \mu_i$. Nga provat e Puasonit numri total *n* është i rastit me shumë se i fiksuar. Në qoftë se marrim një model të Puasonit të kushtëzuar nga *n*, $\{X_i\}$ nuk duhet të ketë shpërndarje të Puasonit, meqënëse çdo X_i nuk mund ta tejkaloj *n*. Me tu dhënë *n*, ndryshoret e rastit $\{X_i\}$ nuk janë të pavarura, pasi vlera e njërit ndikon në një varg të mundshëm vlerash për një ndryshore tjetër. Për *c* ndryshoret e pavarura të Puasonit, me $E(X_i) = \mu_i$, le te marrim shpërndarjet e tyre të kushtëzuara $\sum X_i = n$. Probabiliteti me kusht i një grupi numërimesh $\{n_i\}$ plotëson këtë kusht:

$$P\left(X_1 = n_1, X_2 = n_2, \dots, X_c = n_c \middle| \sum X_j = n\right) = \frac{P(X_1=n_1, X_2=n_2, \dots, X_c=n_c)}{P(\sum X_j = n)} =$$

$$= \frac{\prod_i \left[\frac{\exp(-\mu_i)\mu_i^{n_i}}{n_i!}\right]}{\exp(-\sum \mu_j)(\sum \mu_j)^n/n!} \qquad (1.4)$$

$$= \frac{n!}{\prod_i [n_i!]} \prod_i p_i^{n_i}$$

ku $\{p_i = \mu_i/(\sum \mu_j)\}$. Kjo është një shpërndarje multinomiale e karakterizuar nga vëllimi i zgjedhjes *n* dhe nga probabilitetet $\{p_i\}$. Shumë analiza të të dhënave kategorike merren me shperndarjet multinomiale. Këto analiza zakonisht kanë të njëjtin parametër të vlerësuar si ato të analizave të marra nga një shpërndarje e Puasonit, për shkak të ngjashmërisë në funksionet e përgjasisë maksimale.

1.5 Vlerësimet Satistikore për të dhënat kategorike

Zgjedhja e shpërndarjes probabilitare për ndryshoret e rastit në studim është vetëm një hap i analizës së të dhënave. Në praktikë, shpërndarja ka parametra të panjohur. Në vazhdim trajtohen metodat e vlerësimit të parametrave.

1.5.1 Funksioni i përgjasisë dhe vlerësimet me përgjasinë maksimale

Metoda e përgjasië maksimale përdoret për të vlerësuar parametrat e panjohur. Metoda e përgjasisë maksimale (MPM) është propozuar nga Fisher (1922) dhe ka gjetur përdorime shumë të gjëra sepse jep vlerësues me veti mjaft të mira.

Le të jenë dhënë $X_1, X_2, \dots, X_n$ një zgjedhje nga një popullim Ω ku vrojtohet n.r X, shpërndarja e së cilës njihet, por varet nga parametri i panjohur θ. Vektori i rastit $\vec{X} = (X_1, X_2, \dots, X_n)$ ka shpërndarje :

$\prod_{i=1}^{n} f(x_i, \theta)$ për rastin kur n.r. X është e vazhduar

dhe

$\prod_{i=1}^{n} P(X = x_i)$ për rastin kur n.r. X është diskrete,

ku $f(x, \theta)$ është densiteti i n.r. X. Në qoftë se X_i – njihen dhe $\theta \in \Theta$ është ndryshore atëherë me përkufizim do të themi:

Përkufizim 1.5.1: Funksioni i përgjasisë për parametrin θ, që i korrespondon një zgjedhjes $(X_1, X_2, \dots, X_n)$ quhet funksioni L: Θ → R i përcaktuar me barazimin

$$L\left(\vec{X}, \theta\right) = L(x_1, x_2, \dots, x_n, \theta) = L\left(\vec{X}\right) = \begin{cases} \prod_{i=1}^{n} f(x_i, \theta), \text{për } rastin\ e\ vazhduar \\ \prod_{i=1}^{n} P(X = x_i)\ \text{për } rastin\ diskret \end{cases} \quad (1.5)$$

ku $f(x, \theta)$ është densiteti i n.r. X.

Kështu funksioni i përgjasisë quhet vetë densiteti i zgjedhjes, më në përgjithësi, i vrojtimit, por i shqyrtuar si funksion i parametrit për zgjedhje të fiksuar Naqo (2005). Vëmendja drejtohet tek parametri duke krijuar mundësinë që për një zgjedhje të fiksuar të shihet cila vlerë e parametrit mund të zgjidhet për të plotësuar ndonjë kusht.

Përkufizim 1.5.2: Vlera $\widetilde{\theta}$ e parametrit θ për të cilin:

$$L\left(\widetilde{\theta}\right) = L\left(x_1, x_2, \dots, x_n, \widetilde{\theta}\right) = \sup_{\theta \in \Theta} \{L(x_1, x_2, \dots, x_n, \theta)\}$$

quhet vlerësim i përgjasisë maksimale. Kjo do të thotë se në pikën $\theta = \widetilde{\theta}$ të bashkësisë Θ funksioni i përgjasisë $L(\theta)$, ka maksimum. Ku vlera $\widetilde{\theta}$ gjendet duke zgjidhur ekuacionin:

$$\frac{dL(\theta)}{d\theta} = 0 \text{ ose } \frac{dlnL(\theta)}{d\theta} = 0.$$

(Ku $L(\theta)$ dhe $lnL(\theta)$ kanë të njetin maksimum sepse funksioni ln- është monoton rritës.) Për të vlerësuar r- parametrat $\theta_1, \theta_2, \dots, \theta_r$ duhet të zgjidhet një nga sistemet:

$$\frac{\partial L(\theta_1,\theta_2,\dots,\theta_r)}{\partial \theta_i} = 0 \text{ për } i = 1,2,\dots,r$$

ose

$$\frac{\partial lnL(\theta_1,\theta_2,\dots,\theta_r)}{\partial \theta_i} = 0 \text{ për } i = 1,2,\dots,r.$$

Kjo metodë nuk pohon asgjë për cilësitë e vlerësimit, megjithatë këto vlerësime në përgjithësi janë konvergjente dhe asimptotikisht efikase.

1.5.2 Veti të Vlerësuesit të përgjasisë maksimale (VPM)

Vetia 1: Në qoftë se për parametrin ekziston një statistikë e mjaftueshme, atëherë VPM është funksion i saj.

Vërtetim.

Në bazë të lemës së faktorizimit për statistikën e mjaftueshme shkruajmë:

$$L\left(\vec{X}, \theta\right) = q(\theta, T)r(x).$$

Funksioni L arrin vlerën më të madhe (në lidhje me θ) aty ku arrin funksioni $q(\theta, T)$ dhe kjo pikë del funksion i statistikës së mjaftueshme T.

Vetia 2: Në qoftë se për parametrin ekziston një vlerësues efikas, atëherë VMP është funksion i saj.

Vërtetim.

Vlerësuesi efikas T plotëson barazimin në mosbarazimin Kramer-Roa, pra

$$\frac{\partial(lnp(\underline{x};\theta))}{\partial\theta} \equiv \frac{\partial\left(lnL(\theta;\underline{x})\right)}{\partial\theta} = C(\theta)(T-\theta).$$

Kështu, vlerësuesi efikas T është zgjidhje e ekuacionit të përgjasisë. Derivojmë në lidhje me θ:

$$\frac{\partial^2 lnL}{\partial \theta^2} = \frac{dC}{d\theta}(T-\theta) - C(\theta).$$

Po të zëvendësojmë këtu zgjidhjen e ekuacionit të përgjasisë që është T, gjejmë:

$$\frac{\partial^2 lnL}{\partial \theta^2} = -C(T) < 0,$$

sepse $E\left(\frac{\partial^2 lnL}{\partial \theta^2}\right) = -E((\frac{\partial(lnL)}{\partial \theta})^2) < 0,\ për\ cdo\ \theta.$

Kjo tregon se aty arrihet maksimumi.

Vetia 3: VPM i funksionit $g(\theta)$, g: $R^K \to R^q$ quhet statistika $g(\widetilde{\theta})$, ku $\widetilde{\theta}$ është VPM(θ) d.m.th. VPM($g(\theta)$) = g(VPM(θ)).

Kjo veti është quajtur *invariancë.*

1.5.3 Funksioni i përgjasisë për parametrin binomial

Pjesa e një funksioni të përgjasisë që përmban parametrat quhet *kernel*. Meqënëse maksimizojmë funksionin e përgjasisë lidhur me parametrat, pjesa e mbetur është e parëndësishme. Për këtë marrim shpërndarjen binomiale (1.1) dhe funksionin e përgjasisë (1.5).

Koeficenti binomial C_n^x nuk ka ndikim aty ku maksimumi shfaqet në lidhje me *p*. Kështu, që ai nuk merret parasysh dhe trajtohet si *kernel*, funksioni i përgjasisë. Funksioi i përgjasisë logaritmike për shpërndarjen binomiale është:

$L(p) = \log[p^x(1-p)^{n-x}] = xlog(p) + n - x)\log(1-p).$ (1.6)

Derivojmë në lidhje me p dhe e barazojmë me zero:

$\partial L(P)/\partial p = x/p - (n-x)/(1-p) = (x-np)/p(1-p)$ (1.7)

$\partial L(P)/\partial p = 0$

Marrim $\hat{p} = x/n$, për *n* prova. Ku x është numri i sukseseve. Llogarisim $\partial^2 L(P)/\partial p^2$, duke marrë dhe pritjen arrijmë në këtë rezultat:

$$-E(\partial^2 L(P)/\partial p^2) = E[x/p^2 + (n-x)/(1-p)^2] = n/[p(1-p)].$$

Kështu varianca asimptotike për $\hat{p}$ është $p(1-p)/n$.

Meqenëse $E(X) = np$ dhe $Var(X) = np(1-p)$ atëherë shpërndarja $\hat{p} = X/n$ ka pritje matematike dhe devijim standart

$E(\hat{p}) = p, \sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$.

1.5.4 Testi për një parametër binomial

Për hipotezën $H_0: p = p_0$, përdoret statistika e Wald-it:

$$z_w = \frac{\hat{p}-p_0}{SE} = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p}/n}} \quad (1.8)$$

Ku SE është gabimi standart. Duke vlerësuar rezultatin binomial (1.7) dhe informacionet (1.8) për p_0 merret:

$u(p_0) = \frac{x}{p_0} - \frac{n-x}{1-p_0}$, $L(p_0) = \frac{n}{p_0(1-p_0)}$

Trajta normale e rezultatit statistikor thjeshtohet në statistikën (1.9)

$$z_S = \frac{u(p_0)}{L(p_0)^{1/2}} = \frac{x-np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \quad (1.9)$$

Ndërsa statistika e Wald-it z_W përdor gabimin standart për të vleresuar $\hat{p}$, statistika e rezultateve z_S e përdor atë për të vlerësuar p_0. Përdorim gabimin standart SE real në vend të atij të vlerësuar. Shpërndarja e zgjedhjes është më afër me atë standarte normale sesa statistika e Wald-it. Logaritmi i funksionit të përgjasisë (1.9) është i barabartë me:

$$L_0 = xlogp_0 + (n-x)\log(1-p_0) \text{ për } H_0$$

dhe

$$L_1 = xlog\,\hat{p} + (n-x)\log(1-\hat{p}) \text{ për } H_a$$

Testi statistikor i përgjasisë thjeshtohet në

$$-2(L_0 - L_1) = 2(xlog\frac{\hat{p}}{p_0} + (n-x)log\frac{1-\hat{p}}{1-p_0})$$

I shprehur si

$$-2(L_0 - L_1) = 2(xlog\frac{x}{np_0} + (n-x)log\frac{n-x}{n-np_0})$$

I cili krahason sukseset dhe mossukseset të llogaritur me

$$2\sum e\ vëzhguar\ log\frac{e\ vëzhguar}{e\ përshtatur} \qquad (1.10)$$

Vemë re se kjo formulë përmban teste për parametrat e Puasonit dhe multinomiale. Meqenëse asnjë parametër i panjohur nuk nevojitet për H_0 dhe një nevojitet për H_a, (1.10) ka shpërndarje hi-katror asimptotike me shkallë lirie $df = 1$.

1.5.5 Intervalet e besimit për një parametër

Rëndësia e një testi tregon nëse një vlerë e veçantë p (për shembull p = 0.5) është e besueshme. Përdorim intervalin e besimit për të përcaktuar zonën e vlerave të besimit. Testi i Wald-it jep intervalin e vlerave p_0 për të cilën

$$|z_W| < z_{\alpha/2}$$

ose

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p}/n} \qquad (1.11)$$

Historikisht, ky është një nga intervalet e parë të besimit që është përdorur për çdo parametër. Ai jep rezulatate të gabuara kur *n* është shumë e madhe. Një model i thjeshtë që i shton $\frac{1}{2}z^2_{\alpha/2}$ vrojtimeve të çdo tipi të zgjedhjes përpara se të përdori këto formula ka një rezultat me të mirë. Intervali i besimit të rezultateve përmban vlerën p_0 për të cilën $|z_S| < z_{\alpha/2}$.

Përfundimet e tij janë zgjidhjet p_0 të ekuacioneve $(\hat{p}-p_0)\sqrt{\hat{p}(1-\hat{p}/n} = \pm z_{\alpha/2}$.

Këto janë ekuacione të gradës së dytë për p_0.

Ky interval është:

$$\hat{p}\left(\frac{n}{n+z_{\alpha/2}{}^2}\right)+\frac{1}{2}\left(\frac{z_{\alpha/2}{}^2}{n+z_{\alpha/2}{}^2}\right)$$
$$\pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}{}^2}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z_{\alpha/2}{}^2}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right]\left(\frac{z_{\alpha/2}{}^2}{n+z_{\alpha/2}{}^2}\right)}$$

Termi $\hat{p}$ i intervalit është pesha mesatare e $\hat{p}$ dhe ½ ës, ku pesha e dhënë $\left(\frac{n}{n+z^2_{\alpha/2}}\right)$ çon në rritjen e $\hat{p}$ me rritjen e *n*. Duke i kombinuar termat marrim barazimin e mëposhtëm:

$$\check{p} = \left(x + z^2_{\alpha/2}/2\right)/(n + z^2_{\alpha/2})$$

Ky është një interval për një model të përmirësuar i cili i shton vrojtimeve $z^2_{\alpha/2}$, gjysmën për çdo tip. Katrori i koefiçentit $z_{\alpha/2}$ në këtë formulë është pesha mesatare e dispersionit të një zgjedhje kur $p = \hat{p}$ dhe dispersioni i modelit kur $p = \frac{1}{2}$, duke përdorur vëllimin e zgjedhjes $n + z^2_{\alpha/2}$ në vend të n. Ky interval ka një efektshmëri më të mirë se sa intervali i Wald-it. Intervali i besimit të përgjasisë është i ndërlikuar në llogaritje por i thjeshtë në parim. Ai është një bashkësi p_0 për të cilin testi i përgjasisë ka një p-vlerë shumë herë më të madhe se α. E barasvlefshme me një bashkësi p_0 për të cilën përgjasia logaritmike zvogëlohet më pak se

$\chi^2_1(\alpha)$ nga vlera e vlerësuar e përgjasisë maksimale $\hat{p} = x/n$.

Shembull 1.1.2

Në një grup studentesh me 25 vete u bë pyetja A je vegjetarian? Për n=25 studentët e pyetur u shënua me y = 0 përgjigja "po". Kjo nuk është zgjedhje rasti nga një popullim i caktuar, por po i përdorim këto të dhëna për të ndërtuar një interval besimi me 95% siguri me parameter binomial p.

Nëse y=0, $\hat{p}$ =0/25=0. Duke përdorur intervalin e Wald-it, intervali i besimit 95% për p është:

$$0 \pm 1.96\sqrt{(0.0 \times 1.0)/25}$$

Në qoftë se vrojtimet bien në kufi me hapësirën e zgjedhjes, përdorimi i metodës së Wald-it nuk siguron zgjidhje të saktë. Intervali i besimit me 95% siguri eshte i barabartë me (0.0, 0.133).

Për hipotezën $H_0: p = 0.5$ për shembull testi statistikor është:

$$z_S = (0 - 0.5)/\sqrt{(0.5 \times 1.0)/25} = -5.0$$

Duke e krahasuar me hipotezën $H_a: p = 0.10, z_S = (0 - 0.10)/\sqrt{(0.10 \times 0.90)/25} = -1.67$. Kur y = 0 dhe n = 25, kernel-i i funksionit te pergjasise është $l(p) = p^0(1-p)^{25} = (1-p)^{25}$. Logartimi i përgjasisë (1.8) është L(p) =25 log(1-p).

Vëmë re se: L($\hat{p}$) $= L(0) = 0$.

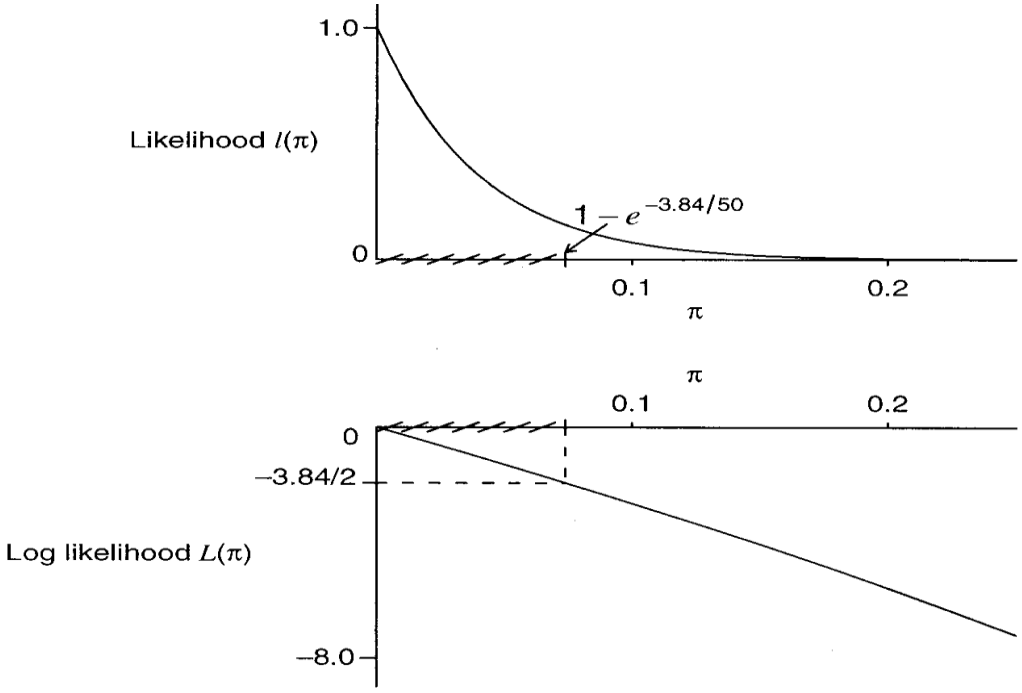
Intervali i besimit me 95 % siguri për p është një grup p_0 për të cilën statistika e përgjasisë është: $-2(L_0 - L_1) = -2[L(p_0) - L(\hat{p})]$

$= 50\log(1 - p_0) \leq \chi^2_1(0.05)$ = 3.84.

Kufiri i sipërm është $1 - \exp(-3.84/50) = 0.074$, dhe intervali i besimit është i barabartë me]0.0, 0.074[.

Figura 1.1 paraqiten funksioni i përgjasisë dhe logaritmi i funksionit të përgjasisë kur y=0 për n=25 prova, dhe interval besimi për p. Tre metodat për zgjedhje të mëdha prodhojnë rezultate të ndryshme. Kur *p* është afër 0, shpërndarja e zgjedhjes $\hat{p}$ është shumë e pjerrët nga e djathta për *n* te vogla. Ajo konsiderohet si nje metodë alternative jo sipas një vleresimi asimptotik.

Figura 1.1: Funksioni i përgjasisë dhe logaritmi i funksionit të përgjasisë



1.6 Vlerësimet statistikore për parametrat multinomial

Në këtë pjesë do të paraqesim vlerësimet për parametrat multinomial $\{p_j\}$. Në n vrojtime, n_j ndodhin në kategorinë $j, j = 1,2, \dots, c$.

1.6.1 Vlerësimet për parametrat multinomial

Me anë të metodës së përgjasisë maksimale vlerësojmë $\{p_j\}$. Probabiliteti multinomial është propocional me Kernel

$$\prod_j p_j^{n_j}$$

ku $p_j \geq 0$ dhe $\sum_j p_j = 1$.

Funksioni logaritmik i përgjasisë multinomiale është

$$L(p) = \sum_j n_j \, logp_j$$

Duke larguar tepricat, atëherë L është funksion i $(p_1, p_2, \dots, p_{c-1})$ ku marrim që:

$$p_c = 1 - (p_1 + p_2 + \cdots + p_{c-1}), \partial p_c / \partial p_j = -1, j = 1,2, \dots, c-1.$$

Duke derivuar $logp_c$ në lidhje me p_j kemi:

$$\frac{\partial logp_c}{\partial p_j} = \frac{1}{p_c}\frac{\partial p_c}{\partial p_j} = -\frac{1}{p_c}$$

Derivojmë $L(p)$ në lidhje me p_j marrim ekuacionin e përgjasisë si më poshtë:

$$\frac{\partial L(p)}{\partial p_j} = \frac{n_j}{p_j} - \frac{n_c}{p_c} = 0, \quad j = 1, \dots, c-1.$$

Nga metoda e përgjasisë maksimale marrim statistikën;

$$\hat{p}_j \Big/ \hat{p}_c = {}^{n_j}\!/_{n_c}, \sum_j \hat{p}_j = 1 = \frac{\hat{p}_c\left(\sum_j n_j\right)}{n_c} = \frac{\hat{p}_c n}{n_c},$$

kështu $\hat{p}_c = n_c/n$ marrim $\hat{p}_j = {}^{n_j}\!/_{n}$

1.6.2 Testi për parametrin Multinomial

Përpara ndërtimit të teorisë së përgjithshme të kontrollit të hipotezave, K. Pearson (1990) propozoi krahasimin e vlerave të vrojtuara V_i me vlerat e pritshme T_i. Caktohet një masë për të matur mënjanimin e vlerave të vrojtuara nga vlerat teorike dhe një vlerë kritike e këtij mënjanimi: në qoftë se mënjanimi faktik e kalon këtë vlerë kritike, atëherë hipoteza për barazimin e probabiliteteve me vlerat hipotetike hidhet poshtë. Kështu konsiderohet

$H_0: p_j = p_{j0}, j = 1,2, \dots, c$ ku $p_{j0} \geq 0, \ \sum_j p_{j0} = 1.$

Për të kontrolluar këtë hipotezë përdoret statistika e Pearsonit: $\chi^2 = \sum_j \frac{(V_i - T_i)^2}{T_i} = \sum_j \frac{(n_j - np_{j0})^2}{np_{j0}}$,

Kur është e vërtetë hipoteza zero dhe $n \to \infty$, ka shpërndarje asimptotike hi-katror me $(k-1)$ shkallë lirie: $\chi^2 \underset{n \to \infty}{\overset{H_0}{\sim}} \chi^2(k-1)$. Zona kritike për nivelin α është: $\chi^2 > \chi^2_{1-\alpha}(k-1)$.

1.7 Metadat statistikore Bayes-iane për të dhënat kategorike dhe teorema e Bayes-it

Në teoremën Bayes-iane me vektorin e parametrave, informacioni aktual i përftuar nga të dhënat Y me densitet $f(Y/\boldsymbol{\beta})$ kombinohet me informacionin paraprak në mënyrë që shpërndarja paraprake (prior) e vlerave të parametrave të panjohur me densitet $P(\boldsymbol{\beta})$ rezulton në shpërndarjen të mëpasshme (posterior) $P(\boldsymbol{\beta}/Y)$ përmes kësaj teoreme:
Teorema e Bayes

$$\pi(\boldsymbol{\beta}/\mathrm{Y}) = \frac{\mathrm{f}(\mathrm{Y}/\boldsymbol{\beta})\mathrm{P}(\boldsymbol{\beta})}{\int \mathrm{f}(\mathrm{Y}/\boldsymbol{\beta})\mathrm{P}(\boldsymbol{\beta})\mathrm{d}(\boldsymbol{\beta})} \qquad (1.12)$$

ku $m(x) = \int \mathrm{f}(\mathrm{Y}/\boldsymbol{\beta})\mathrm{P}(\boldsymbol{\beta})\mathrm{d}(\boldsymbol{\beta})$ është densiteti me kusht i Y ose probabiliteti i të dhënave për modelin me parametër $\boldsymbol{\beta}$. Densiteti i Y është një problem i vështirë për tu zgjidhur, kështu që në mënyrë të ngjashme përdoret shpesh formula:

$$\pi(\boldsymbol{\beta}/\mathrm{Y}) \approx \mathrm{f}(\mathrm{Y}/\boldsymbol{\beta})\,\mathrm{P}(\boldsymbol{\beta}) \qquad (1.13)$$

Në këtë ekuacion shpërndarja posterior është në proporcion me produktin e mëparshëm dhe me probabilitetin.
Peshat relative të mundësisë dhe priori përcaktohen nga varianca e shpërndarjes sipas Simon et al. (2010). Një shpërndarje me variancë të vogël paraqet një peshë të lartë në përcaktimin e posteriorit. Për shembull:
Një shpërndarje normale prior $P(b_1) \sim N(\mu = 10, \sigma = 10)$ ka më pak ndikim në mundësinë e ndodhjes sesa një prior me shpërndarje normale $P(b_2) \sim N(\mu = 10, \sigma = 2)$.

Teorema e Bayes-it pohon se informacioni i përdorur në probabilitetet e përftuara rreth ngjarjeve të vrojtuara subjektive jep vlerësim.
Për shembull në një model linear të thjeshtë regresi $\hat{y} = b_0 + b_1 x$ nëse njohim dicka rreth vlerës së pritshme (mesatares) të parametrit të panjohur b_1 të modelit nga kërkimet e mëparshme, nga njohuritë themelore ose nga të dhënat prior, mund të pohojmë se $b_1 \sim N(\mu, \sigma)$ ku μ dhe σ janë parametra të njohur të një shpërndarje normale. Për vlera të mëdha të σ mund të përftojmë shkallë të lartë pasigurie për vlerën e b_1, ndërsa vlera të vogla mund të përftojmë besim të lartë. Në mënyrë alternative mund të pohojmë se $b_1 \sim U(l, u)$ ku l dhe u janë kufiri i poshtëm dhe kufiri i sipërm i një shpërndarje uniforme të vazhdueshme.

Për modelet statistikore Bayes-iane, me rëndësi janë dhe avantazhet dhe disavantazhet e tyre. Një kritikë e fortë për këto modele është selektimi i shpërdarjes prior që ndikon në vlerën e parametrit. Për shembull, supozojmë se 10 studime në një fenomen gjatë një vrojtimi tregon bashkësinë e vlerave {5.5, 8.1, 9.0, 6.3, 5.3, 6.5, 3.9} për parametrin b_1 në një model të thjeshtë të regresit linearë.
Si mund ta paraqesim këtë informacion në një prior?
Mund të përafrojmë këto të dhëna me një shpërndarje normale dhe të përdorim parametrat e përftuar si prior. Në qoftë se këto vlera nuk dominojnë parametrat e modelit, atëherë varianca do të rritet duke pasqyruar besueshmëri të lartë ose të ulët në këto vlera. Sido të jetë zgjedhja e një priori është e justifikueshme nga disa studiues dhe shpesh është një sfidë për të argumentuar bindshëm rreth një priori mbi të tjerët.
Sic u përmend më lart, një supozim i përgjithshëm i metodave Bayes-iane është se asnjë informues nuk është i njohur si prior duke lejuar në këtë mënyrë përdorimin e priorëve jo informues dhe më tej largimin e tyre. Për shembull duke supozuar se një parametër ka pritje matematike zero dhe shmagie mesatare 1000 – relativisht një prior jo informues. Sidoqoftë, caktimi i priorve jo informues si në këtë shembull, shpesh rezulton në një model me parametra që nuk mund të përftohen nëpërmjet metodës së përgjasisë maksimale.

1.7.1 Shpërndarja prior për një parametër Binomial

Le të jetë Y ndryshore rasti me shpërndarje Binomiale me parametra n dhe π, shënojmë $p = {}^{y}/_{n}$
Meqë $\pi \in (0,1)$, sipas Agresti et al. (2005) një shpërndarje prior për π është shpërndarja beta me parametra $\alpha > 0, \beta > 0$ e cila është proporcional me:

$$\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

Kjo shpërndarje ka $E(\pi) = {}^{\alpha}/_{\alpha+\beta}$.
Densiteti posterior $k(\pi/y)$ i π është proporcional me

$$k(\pi/y) \sim [\pi^{y}(1-\pi)^{n-y}]\left[\pi^{\alpha-1}(1-\pi)^{\beta-1}\right] =$$

$$= \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1},$$

për $0 \le \pi \le 1$ dhe me shpërndarje beta. Beta është shpërndarja prior. Shpërndarja posterior është gjithashtu beta. Si rrjedhojë konkludojmë:

- π ka shpërndarje beta me parametra $\alpha^* = y + \alpha$ dhe $\beta^* = n - y + \beta$. Në mënyrë ekuivalente, kjo është shpërndarja e $\frac{\left(\frac{y+\alpha}{n-y+\beta}\right)F}{1+\left(\frac{y+\alpha}{n-y+\beta}\right)F}$.

ku F është n.r me shpërndarje Fisheri me $n_1 = 2(y + \alpha)$ dhe $n_2 = 2(n - y + \beta)$ shkallë lirie.

- $\left(\frac{n-y+\beta}{y+\alpha}\right)\frac{\pi}{1+\pi}$ ka shpërndarje Fisheri me n_1=2(y+α) dhe n_2=2(n-y+β) shkallë lirie.

Mesatarja e shpërndarjes posterior beta të π është një mesatare e peshave e proporcionit të zgjedhjes dhe mesatares së shpërndarjes prior

$$E(\pi/y) = {\alpha^*}/{\alpha^* + \beta^*} = {(y + \alpha)}/{(n - y + \beta)}$$

$$= w\left(\frac{y}{n}\right) + (1 - w)\left[{\alpha}/{(\alpha + \beta)}\right]$$

ku $w = {n}/{(n - y + \beta)}$ është mesatarje e peshave të zgjedhjes.

Varianca e shpërndarjes posterior është

$$Var(\pi/y) = {\alpha^*\beta^*}/{(\alpha^* + \beta^*)^2(\alpha^* + \beta^* + 1)}$$

që përafrohet me $\sqrt{{p(1-p)}/{n}}$ për n të mëdha.

Vlerësuesi i përgjasisë maksimale $p = {y}/{n}$ rezulton për $\alpha = \beta = 0$.

Për shpërndarjen uniforme prior $\alpha = \beta = 1$, shpërndarja posterior ka të njetën përmasë me funksionin e përgjasisë maksimale binomial me pritje matematike $E(\pi/y) = (y + 1)/(n + 2)$.

Përvec shpërndarjes uniforme, priori më i përdorur i vlerësimit Binomial është priori Jeffrey. Ky prior është proporcional me rrënjën katrore të përcaktorit të matricës informuese Fisher për parametrat e interesuar, gjatë një vëzhgimi. Në rastin e shpërndarjes Binomiale ky prior është beta me $\alpha = \beta = 0.5$.

Një vlerësim altenativ me dy parametra specifikon një prior normal për $logit(\pi)$. Kjo shpërndarje e π është quajtur logistic-normal. Me shpërndarje prior$N(0, \sigma^2)$ për $logit(\pi)$ densiteti prior për π është:

$$f(\pi) = \frac{1}{\sqrt{2(3.14)\sigma^2}} \exp\{\frac{-1}{2\sigma^2}(log\frac{\pi}{1-\pi})^2\}\frac{1}{\pi(1-\pi)}, 0 < \pi < 1$$

Në hapësirën probablitare (π) ky densitet është simetrik, bëhet unimodal kur $\sigma^2 \leq 2$ dhe bimodal kur $\sigma^2 > 2$, por gjithmonë tenton drejt zeros ashtu si π merr vlerën zero dhe një.

- Për $\sigma = 1$ densiteti grafikisht paraqitet me vlera në rritje graduale dhe më pas këto vlera fillojnë e zbresin.
- Për $\sigma = 0.5$ densiteti është pothuajse uniforme bën përjashtim sjellja e tij afër kufijve.
- Për $\sigma = 2$ densiteti grafikisht paraqitet me vlera në rritje që arrijnë vlerën maksimale dhe më pas zbresin.
- Për $\sigma = 3$ sjellja e densitetit është e ngjashme me priorin me shpërnarje beta (0.5, 0.5).

Me priorin logjistik-normal, densiteti i posteriorit njehsohet si një integral për konstanten e normalizuar.

1.7.2 Vlerësimi Bayes-ian për parametrin binomial

Përgjithësisht, përdorimi i priorit uniform sjell shpërndarje posterior në ndërtimin e një intervali besimi për një parametër binomial. Për shembull Walters (1984) ka treguar se kufiri i poshtëm π_L i intervalit të besimit me $\gamma = 0.95$ të Clopper–Pearson gëzon barazimin $P(Y \geq y/\pi_L) = 0.25$. Brawn Cai dhe Das Gupa (2001, 2002) kanë treguar se shpërndarja posterior e gjeneruar nga priori Jeffrey për intervalin e besimit të π në termat e mesatares siguron probabilitet dhe supozimet e pritur.

Për hipotezat

$$\boldsymbol{H_0: \pi \geq \pi_0}$$
$$\boldsymbol{H_a: \pi < \pi_0}\ ,$$

një vlerë Bayes-iane P është probabiliteti posterior ku $(\pi \geq \pi_0/y)$. Routledge (1994) tregoi se me prior Jeffreys dhe $\pi_0 = 0.5$ është afërsisht i barabartë me mesataren e njëanshme të vlerës P për testin binomial. Pjesa më e madhe e literaturave rreth vlerësimit Bayes-ian për një parametër binomial merren me rezultatin teorik të vendimit. Për vlerësimin e një parametri θ duke përdorur vlerësuesin **T** me funksion $\omega(\theta)(\boldsymbol{T} - \theta)^2$, vlerësimi Bayes-ian është:

$$E[\theta\omega(\theta)/y]/E[\omega(\theta)/y].$$

Me funksion $[(T - \pi)^2]/\pi(1 - \pi)$ dhe shpërndarjen prior uniforme, vlerësimi Bayes-ian i π merret nga vlerësuesi i përgjasisë maksimale $\check{p} = {}^{y}/_{n}$.

1.7.3 Vlerësimi i parametrave multinomial

Rezultatet për parametrin binomial me shpërndarje prior beta përgjithësohen dhe në ato multinomial me prior Dirichle. Me c-kategori, supozojmjë se qelizat $(n_1, n_2, \dots, n_c)$ kanë shpërndarje multinomiale me $n = \sum n_i$ dhe parametra $\pi = (\pi_1, \pi_2, \dots, \pi_c)'$.

Funksioni i përgjasisë multinomial është proporcional me

$$\prod_{i=1}^{c} \pi_i^{n_i}.$$

Densiteti i konjuguar është Dirichle, i shprehur në termat e funksionit gamma jepet si më poshtë:

$$g(\pi) = \frac{\Upsilon(\sum \alpha_i)}{\prod_i \Upsilon(\alpha_i)} \prod_{i=1}^{c} \pi_i^{\beta_i - 1}$$

Për $0 \le \pi_i \le 1$ për cdo i dhe $\sum_i \pi_i = 1$, ku $\{\alpha_i \ge 0\}$.

Le të shënojmë me $\sum_i \alpha_i = K$, shpërndarja Dirichle ka pritje matematike dhe Variancë si më poshtë:

$E(\pi_i) = {}^{\alpha_i}/_{K}$ dhe $Var\ (\pi_i) = \alpha_i(K - \alpha_i)/K^2(K + 1)$.

Densiteti posterior është gjithashtu Dirichle me parametra $\{n_i + \alpha_i\}$, kështu pritja matematike posterior është:

$$E(\pi_i / n_1, n_2, \dots, n_c) = (n_i + \alpha_i)/n + K$$

Le të jetë $\gamma_i = E(\pi_i) = \alpha_i / K$, ky vlerësues Bayes-ian ekuivalenton mesataren e peshave të mëposhtme:

$$[n/(n + k)]p_i + [k/(n + k)\ ]\gamma_i,$$

e cila është pjesa e modelit kur informacioni prior korespondon me provat K të rezultateve $\alpha_i,\ i = 1, 2, \dots, c$.

Sipas Goog (1965) i referohet K si një konstante, me sjellje të vecantë pasi me të njetat $\{\alpha_i\}$ ky vlerësim përjashton pjesë të modelit ndaj vlerës probabilitare $\gamma_i = \frac{1}{c}$, për n të fiksuar, vlerat e K

rriten. Ndërsa në (1980) Good atribuoi $\{\alpha_i = 1\}$ tek De Morgam (1847), kur përdori $(n_i + 1)\Big/(n + c)$, për të vlerësuar π_i të zgjeruar në vlerësimin e Laplasit në rastin multinomial.

Perkos (1947) sugjeroi $\{\alpha_i = \frac{1}{c}\}$, me prior Jeffreys për parametrat, Hoadly (1969) përdori vlerësimin Bayes-ian të rastit multinomial kur popullimi i interesuar është i fundëm me përmasë të njohur N. Ai argumenton se një popullim i fundëm me prior Dirichle është një prior multinomial kompleks i cili çon tek një posterior multinomial kompleks.

Shënojmë N një vektor me numra të plotë jo negativë, të tillë që komponenti i i-të i N_i është numri i objekteve me kategori $i, i = 1, 2, \dots, c$. Nëse vendosim kushte në probabilitete dhe N që, qelizat e numërueshme kanë shpërndarje multinomiale dhe nëse probabilitetet e tyre multinomiale kanë shpërndarje Dirichle të indeksuar nga parametri α i tillë që $\alpha_i > 0$ për të gjitha i me $K = \sum \alpha_i$, atëherë në mënyrë të pakushtëzuar N ka funksionin e masës multinomial kompleks:

$$f(N/N; \alpha) = \frac{N!\,\Gamma(K)}{\Gamma(N + K)} \cdot \prod_{i=1}^{c} \frac{\Gamma(N_i + \alpha_i)}{N_i!\,\Gamma(\alpha_i)}$$

Ky funksion shërben si një shpërndarje prior për N. Për të dhënat e numërueshme $\{n_i\}$ në një zgjedhje me përmasë n, shpërndarja posterior $N - n$ është multinomiale kur

- N zëvenësohet nga $N - n$ dhe
- α zëvendësohet nga $\alpha + n$.

Për shpërndarjen Dirichle, mund të specifikojmë mesataren duke zgjedhur $\{\gamma_i\}$ dhe variancën nëpërmjet K, por nuk ka liri zgjedhje për korrelacionet. Si një altenativë, propozohet përdorimi I një shpërndarje normale prior me multivariacionin për parametrat multinomial $\leftrightarrow logit$.

Në qoftë se $X = (X_1, X_2, \dots, X_c)$ ka shpërndarje normale me multivariacion atëherë $\pi = (\pi_1, \pi_2, \dots, \pi_c)$ me π_i si më poshtë:

$$\pi_i = \exp(X_i)\Big/\sum_{j=1}^{c} \exp(X_j)$$

ka shpërndarje logjistike $\leftrightarrow$ normale. Për shembull kur kategoritë janë të renditura dhe njëra shfaq ngjashmëri probabilitare në kategorinë e afërt, njëra mund të përdorë një formë autoregresive për matricën e korrelacionit. Leonard (1973) ka sugjeruar këtë përafrim në vlerësimin e një histograme. Dickey (1983) paraqet shpërndarje të cilat gjenerojnë shpërndarjen Dirichle dhe argumenton se ato janë të përshtatshme për tabelat e kontigjencës. Në qoftë se $\pi_1 \le \pi_2 \le \cdots \le \pi_k \ge \pi_{k+1} \ge \cdots \ge \pi_c$, vlerësimi i probabiliteteve multinomiale gjendet duke përdorur priorin e kufizuar Dirichle dhe mundësisht një prior në k në qoftë se është i panjohur.

1.8 Përfundime

Në kapitullin 1 ndryshoret kategorike klasifikohen dhe analizohen së bashku me shpërndarjet probabilitare të tyre. Për shkak të mbidispersionit analiza e shpërndarjes binomiale dhe multinomiale është shpesh e pavlefshme. Në testin për një parametër binomial përdoret gabimi standart SE real në vend të atij të vlerësuar.
Priori më i përdorur i vlerësimit për parametrin binomial është priori Jeffrey, rezultatet për parametrin binomial me shpërndarje prior beta përgjithësohen në ato multinomiale me prior Dirichle. Modelet Bayes-iane që hasen në analizën e të dhënave kategorike përdorin informacionin paraprak (prior) në vlerat e parametrave të të dhënave të vrojtuara në mënyrë që të përftohen vlerësues të mëpasshëm (posterior) dhe tregohet se peshat relative të mundësisë dhe priori përcaktohen nga varianca e shpërndarjes.

Kapitulli 2

TABELAT E KONTIGJENCËS DHE STRUKTURA PROBABILITARE E TYRE

2.1 Hyrje

Në këtë kapitull trajtohen tabelat dhe matricat e kontingjencës me dy-hyrje. Tabelat e kontingjencës me 2-hyrje mund të zgjerohen në tabela të kontingjencës me shumë hyrje për ndryshoret multinomiale. Ndërtohet matrica e kontigjencës, elementët e së cilës janë të barabarta me vlerat korresponduese të tabelës së kontigjencës duke përjashtuar vlerat margjinale. Tregohet sipas Tsumoto (2009) se kur një tabelë kontigjence merret si matricë atëherë për pavarsinë statistikore midis ndryshoreve rol kryesor luan rangu i matricës. Tregohet se për r modele të pavarura me shpërndarje binomiale, tabela e kontigjencës ka $r \times 2$ përmasë dhe vlerësimi i parametrave binomial kryhet duke përdorur vlerësimin empirik Bayes-ian.
Bëhet vlerësimi i parametrave loglinear në tabelat e kontigjencës me dy hyrje, nga moda e tyre posterior e cila shërben për të përftuar vlerësues të qelizave probabilitare.

2.2 Struktura probabilitare për tabelat e kontigjencës

Shpërndarja probabilitare midis ndryshoreve kategorike përcakton lidhjen e tyre. Ato mund të jenë të varura ose të pavarura.

2.2.1 Tabelat e kontigjencës dhe shpërndarjet e tyre

Përkufizim 2.2.1. Një tabelë statistikore e cila tregon denduritë e vrojtuara të elementeve të të dhënave të klasifikuara të dy ndryshoreve, me rreshtave që tregojnë një ndryshore dhe shtyllave që tregojnë ndryshoren tjetër quhet tabelë e kontigjencës me dy hyrje.

Por Sipas Karl Pearson (1904) kemi këtë përkufizim për tabelat e kontigjencës:

Përkufizim 2.2.2. Qelizat e një tabele që përmbajnë denduritë e llogaritura nga rezultatet e vrojtuara për një zgjedhje, tabela është quajtur *tabelë kontigjence*.

Një emer tjetër është *tabela e klasifikimit te kryqezuar*. Një tabelë kontigjence me *I* rreshta dhe *j* shtylla është quajtur tabelë $i \times j$.

Le të jetë *X* dhe *Y* dy ndryshore kategorike, *X* i kategorisë *k* dhe *Y* i kategorisë *m*. Klasifikimi i të dy ndryshoreve ka *m*×*k* mundësi kombinimi. Ndryshoret *(X,Y)* të zgjedhura në mënyrë të rastësishme nga një popullim kanë një shpërndarje probabilitare. Në një tabelë kemi *m* rreshta për kategorine *Y* dhe *k* shtylla për kategorinë *X* që përcaktojnë këtë shpërndarje. Të dhënat paraqiten në tabelën e kontigjencës (2.1)

Tabela 2.1: Tabela e kontigjencës për dy karakteret A dhe B

A B	A_1	A_2	…..	A_j	…..	A_k	Shuma
B_1	x_{11}	x_{12}	…..	x_{1j}	……	x_{1k}	$x_{1.}$
B_2	x_{21}	x_{22}	…..	x_{2j}	…..	x_{2k}	$x_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_i	x_{i1}	x_{i2}	…..	x_{ij}	…..	x_{ik}	$x_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_m	x_{m1}	x_{m2}	…..	x_{mj}	…..	x_{mk}	$x_{m.}$
Shuma	$x_{.1}$	$x_{.2}$	…..	$x_{.j}$	…..	$x_{.k}$	$x_{..} = n$

x_{ij} – është numri i personave që i përkasin klasës së i- të të karakterit B dhe klasës së *j* –të të karakterit A.

$x_{.j} = \sum_{i=1}^{m} x_{ij},\ x_{i.} = \sum_{j=1}^{n} x_{ij},\ \sum_{j=1}^{n} x_{.j} = \sum_{i=1}^{m} x_{i.} = x_{..} = n$.

2.2.2 Tabelat e kontigjencës 2×2

Tabelat e kontigjencës me 2×2 tregojnë lidhjen midis dy karaktereve sipas dendurive të tyre.

Përkufizim 2.2.3. Le të jenë X_1 dhe X_2 ndryshore binare në hapësirën A të ndryshoreve. A është tabelë kontigjence e një bashkësie të fundme si:

$$|[X_1 = 0]|, |[X_1 = 1]|, |[X_2 = 0]|, |[X_2 = 1]|,$$

$$|[X_1 = 0 \wedge X_2 = 0]|\ |[X_1 = 0 \wedge X_2 = 1]|,$$

$$|[X_1 = 1 \wedge X_2 = 0]|, |[X_1 = 1 \wedge X_2 = 1]|,$$

$$|[X_1 = 0 \vee X_1 = 1]|, |[X_2 = 0 \vee X_2 = 1]|(= N)$$

Kjo tabelë është paraqitur në tabelën 2.2 si më poshtë:

$|[X_1 = 0]| = x_{11} + x_{21} = x_{.1}, |[X_1 = 1]| = x_{12} + x_{22} = x_{.2},$

$|[X_2 = 0]| = x_{11} + x_{12} = x_{1.}, |[X_2 = 1]| = x_{21} + x_{22} = x_{2.},$

$|[X_1 = 0 \wedge X_2 = 0]| = x_{11}\ |[X_1 = 0 \wedge X_2 = 1]| = x_{21},$

$|[X_1 = 1 \wedge X_2 = 0]| = x_{12}, |[X_1 = 1 \wedge X_2 = 1]| = x_{22},$

$|[X_1 = 0 \vee X_1 = 1]|, |[X_2 = 0 \vee X_2 = 1]|$

$|[X_1 = 0 \vee X_1 = 1]| = x_{.1} + x_{.2} = x_{..},$

$|[X_2 = 0 \vee X_2 = 1]| = x_{1.} + x_{2.} = x_{..}(= N)$

Tabela 2.2: Tabela e kontigjencës 2×2

	$X_1 = 0$	$X_1 = 1$	Total
$X_2 = 0$	x_{11}	x_{12}	$x_{1.}$
$X_2 = 1$	x_{21}	x_{22}	$x_{2.}$
Total	$x_{.1}$	$x_{.2}$	$x_{..}$

E rëndësishme është të kontrollohet hipoteza H_0 lidhur me pavarësinë e dy karaktereve X_1 dhe X_2.

Për të vlerësuar diferencën midis dendurisë teorike dhe asaj të vrojtuar përdoret testi i Pearsonit (kriteri hi-katror):

$$V^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(x_{ij} - \frac{x_{i.} \cdot x_{.j}}{x_{..}})^2}{\frac{x_{i.} \cdot x_{.j}}{x_{..}}} \sim \chi^2((k-1)(m-1))$$

ose

$$V^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(x_{ij} - h_{ij}\right)^2}{h_{ij}} \sim \chi^2\big((k-1)(m-1)\big)$$

ku $h_{ij} = \frac{x_{i.} \cdot x_{.j}}{x_{..}}$

Ky kriter nuk përdor vlerat e ndyshores, që vrojtohet, por vetëm denduritë e tyre, kjo lejon që ky kriter të përdoret edhe për vrojtime që merren nga ndryshoret kategorike.

Në qoftë se p_{ij} është probabiliteti që një vrojtim i përket qelizës (X_1, X_2) sipas Agresti (2002) atëherë kur hipoteza H_0 është e vërtetë do të kishim që $p_{ij} = p_{i.} \cdot p_{.j}$si hipotezë alternative është $H_a: p_{ij} \neq p_{i.} \cdot p_{.j}$

Gjendet vlera e testit V^2 dhe merret vendimi në qoftë se $V^2_{\propto -1} \geq \chi^2((k-1)(m-1))$ atëherë hipoteza H_0 hidhet poshtë në të kundërt ajo pranohet. Le të jetë p probabiliteti që (X,Y) ndodhet në rreshtin i dhe kolonën j. Shpërndarja probabilitare $\{p_{ij}\}$ është *shperndarja e perzier* e X dhe Y. i përcaktojmë ato me $p_{i.}$ është shuma e probabiliteteve për ndryshoret e rreshtave dhe $p_{.j}$ është shuma e probabilteteve për ndryshoret e kolonave, ku shenja "·" përcakton shumën e indekseve e cila është

$p_{i.} = \sum_{j=1}^{n} p_{ij}$ dhe $p_{.j} \sum_{i=1}^{m} p_{ij}$

Ku $\sum_i p_{i.=} \sum_j p_{.j} = \sum_i \sum_j p_{ij}$=1.0, domethënë shuma sipas rreshtave e probabiliteteve dhe shuma sipas shtyllave e probabiliteteve është 1. Megjithatë, një kategori e fiksuar (X_1, X_2) ka një shpërndarje probabilitare. Ajo lidhet me studimin se si kjo shpërndarje ndryshon kur ndryshon kategoria e X_1-it. Nëse një subjekt klasifikohet në i rreshta për X_1, $p_{j/i}$ tregon probabilitetin e klasifikimit në j kolona për X_2, $j = 1, \ldots, J$.

Dime se $\sum_j p_{j/i} =1$. Forma probabilitare $\{p_{1/i},..., p_{J/i}\}$ e *shperndarjës* për X_2 në kategorinë i për

X_1. Shpërndarjet probabilitare zgjerohen në koncept me llogaritje të qelizave të tabelave të kontigjencës. Për shembull një zgjedhje e marrë me shpërndarje Puasoni i trajton qelizat $\{Y_{ij}\}$ si ndryshore rasti te pavarura Puasoni me parametra $\{\mu_{ij}\}$. Densitetit probabilitar për rezultatet e mundshme $\{n_{ij}\}$ është prodhim i probabiliteteve të Puasonit P($Y_{ij} = n_{ij}$) për $i \times j$ qeliza:

$$\prod_i \prod_j \exp(-\mu_{ij})^{n_{ij}}_{\mu_{ij}} / n_{ij}!$$

Kur vëllimi i zgjedhjes n është i fiksuar por rreshtat dhe shtyllat nuk janë, përdorim *nje model te zgjedhjes multinomiale*. Qelizat $i \times j$ janë rezultatet të mundshme. Densiteti probabilitar për njehsimin e qelizave ka trajtën:

$$[n!/n_{11}\,! \ldots n_{IJ}!] \prod_i \prod_j {p_{ij}}^{n_{ij}}$$

Shpesh, vrojtimet për një ndryshore Y të varur shfaqen në varësi të një ndryshore X të pavarur. Për thjeshtesi perdorim simbolin $n_i = n_{i.}$ për shumën sipas rreshtave. Supozojmë se n_i vrojtimet

e Y varen nga i dhe X janë të pavarur, secila me shperndarje probabilitare $\{p_{1/i},\ldots,p_{J/i}\}$. Njehsimet $\{n_{ij},\ j=1,\ldots,J\}$ kënaqin kushtin $\sum_j n_{ij}=n_i$.

Në këtë mënyrë kemi formën Multinomiale të mëposhtme:

$$\frac{n_i!}{\prod_j n_{ij}!}\prod_j p_{j/i}^{n_{ij}} \qquad (2.1)$$

Kur zgjedhjet e rastit për X janë të pavarura, funksioni probabilitar i përzier për një bashkësi të dhënash është prodhim i funksionit multinomial (2.1). Kjo quhet *zgjedhje multinomiale e pavarur* ose *prodhim zgjedhjesh multinomiale*.

Zgjedhjet e pavarura multinomiale janë rezultat i këtyre kushteve: Supozojmë se n_{ij} është rezultat i zgjedhjes së pavarur me pritje matematike $\{\mu_{ij}\}$ ose zgjedhje multinomiale për qelizat IJ me probabilitetet $\{p_{ij}=\mu_{ij}/n\}$. Kur X është një ndryshore e pavarur, përmbush kushtet e statistikes referenciale ku $\{n_i=\sum_j n_{ij}\}$ kur vlerat nuk janë të fiksuara. Sipas $\{n_i\}$ qeliza njehsohet $\{n_{ij},$ j=1,…,J$\}$ ka shpërndarje multinomiale (2.1) me probabilitete $p_{j/i}=\mu_{ij}/\mu_{i.},\{j=1,\ldots,J\}$ dhe qeliza llogarit që rreshtat janë të pavarur. Me këtë kusht themi se rreshtat në total janë të fiksuar dhe i analizojmë të dhënat si modele të veçantë të pavarur.

2.2.3 Tabela e kontigjencës m×n

Tabela e kontigjencës me dy hyrje mund të zgjerohet në tabela të kontigjencës me shumë hyrje për ndryshoret multinomiale. Tabela e kontigjencës sipas Tsumoto (2006) me mxn përmasë jepet në trajtën:

Tabela 2.3: Tabela e kontigjencës mxn

A B	A_1	A_2	…..	A_j	…..	A_n	Shuma
B_1	x_{11}	x_{12}	…..	x_{1j}	…….	x_{1n}	$x_{1.}$
B_2	x_{21}	x_{22}	…..	x_{2j}	…..	x_{2n}	$x_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_i	x_{i1}	x_{i2}	…..	x_{ij}	…..	x_{in}	$x_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_m	x_{m1}	x_{m2}	…..	x_{mj}	…..	x_{mn}	$x_{m.}$
Shuma	$x_{.1}$	$x_{.2}$	…..	$x_{.j}$	…..	$x_{.n}$	$x_{..}=N$

Përkufizim 2.2.4. Le të jenë X_1 dhe X_2 dy ndryshore multinomiale në hapësirën A që ka m dhe n vlera, tabelë kontigjence e bashkësisë A është:

$|[X_1=A_j]|$, $|[X_2=B_i]|$, $|[X_1=A_j \wedge X_2=B_i]|$, $|[X_1=A_1 \wedge X_1=A_2 \ldots \wedge X_1=A_m]|$,

$|[X_2=B_1 \wedge X_2=A_2 \ldots \wedge X_2=A_n]| = (N)$ (i=1,2,…,n,m=1,2,…,m), N= x..

Kjo tabelë është paraqitur në tabelën 2.3 si më lart ku;

$$\left|\left[X_1 = A_j\right]\right| = \sum_{i=1}^{m} x_{1i} = x_{.j}, |[X_2 = B_i]| = \sum_{j=1}^{n} x_{j1} = x_{i.} \left|\left[X_1 = A_j \wedge X_2 = B_i\right]\right| = x_{ij},$$

$$x_{..} = N(i = 1,2, \ldots, n \; dhe \; j = 1,2, \ldots, m).$$

2.2.4 Pavarësia statistikore e tabelave të kontigjencës 2x2

Konsiderojmë tabelën 2.2 të kontigjencës me dy hyrje. Pavarësia statistikore midis X_1 dhe X_2 është:

$$P([X_1 = 0, X_2 = 0]) = P([X_1 = 0]) \cdot P([X_2 = 0])$$

$$P([X_1 = 0, X_2 = 1]) = P([X_1 = 0]) \cdot P([X_2 = 1])$$

$$P([X_1 = 1, X_2 = 0]) = P([X_1 = 1]) \cdot P([X_2 = 0])$$

$$P([X_1 = 1, X_2 = 1]) = P([X_1 = 1]) \cdot P([X_2 = 1]).$$

Meqenëse çdo probabilitet paraqitet si raport i çdo qelizë me N, ekuacionet e mësipërme janë llogaritur si:

$$\frac{x_{11}}{N} = \frac{x_{11}+x_{12}}{N} \cdot \frac{x_{11}+x_{21}}{N},$$

$$\frac{x_{12}}{N} = \frac{x_{11}+x_{12}}{N} \cdot \frac{x_{12}+x_{22}}{N},$$

$$\frac{x_{21}}{N} = \frac{x_{21}+x_{22}}{N} \cdot \frac{x_{11}+x_{21}}{N},$$

$$\frac{x_{22}}{N} = \frac{x_{21}+x_{22}}{N} \cdot \frac{x_{12}+x_{22}}{N}.$$

Meqenëse $N = \sum_{ij} x_{ij}$, atëherë formulat e mëposhtme meren nga katër barazimet e mësipërme dhe kanë trajtën:

$$x_{11}x_{22} = x_{12}x_{21} \; ose \; \frac{x_{11}x_{22}}{x_{12}x_{21}} = 1$$

Në këto kushte ka vend teorema 2.2.1:

Teorema 2.2.1. Nëse dy ndryshore X_1, X_2 në tabelën e kontigjencës 2×2 janë statistikisht të pavarura, atëherë ka vend ekuacioni:

$$x_{11}x_{22} = x_{12}x_{21} \; ose \; \frac{x_{11}x_{22}}{x_{12}x_{21}} = 1$$

Vemë re se barazimi i mësipërm është përcaktori i matricës korresponduese të tabelës së kontigjencës 2×2 përmasa përcaktori i së cilës është i barabartë me zero. Gjithashtu, kur këto katër vlera nuk janë të barabarta me 0, ekuacioni i mësipërm mund të shndërrohet në trajtën e mëposhtme:

$$\frac{x_{11}}{x_{21}} = \frac{x_{12}}{x_{22}}$$

Supozohet se raporti i mësipërm është i barabartë me C, konstante. Atëherë

$$\begin{cases} x_{11} = Cx_{21} \\ x_{12} = Cx_{22} \end{cases}$$

Duke bërë zëvendësime merret:

$$\frac{x_{11} + x_{12}}{x_{21} + x_{22}} = \frac{C(x_{21} + x_{22})}{x_{21} + x_{22}} = C = \frac{x_{11}}{x_{21}} = \frac{x_{12}}{x_{22}}$$

2.2.5 Pavarësia statistikore e tabelave të kontigjencës 2x3

Konsiderohet tabela e kontigjencës 2×3 e paraqitur në tabelën

Tabela 2.4: Tabela e kontigjencës 2×3.

	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$	Shuma
$X_2 = 0$	x_{11}	x_{12}	x_{13}	$x_{1.}$
$X_2 = 1$	x_{21}	x_{22}	x_{23}	$x_{2.}$
Shuma	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{..}$

Pavarësi statistikore midis X_1, X_2 është:

$$P([X_1 = 0, X_2 = 0]) = P([X_1 = 0]) \cdot P([X_2 = 0])$$

$$P([X_1 = 0, X_2 = 1]) = P([X_1 = 0]) \cdot P([X_2 = 1])$$

$$P([X_1 = 0, X_2 = 2]) = P([X_1 = 0]) \cdot P([X_2 = 2])$$

$$P([X_1 = 1, X_2 = 0]) = P([X_1 = 1]) \cdot P([X_2 = 0])$$

$$P([X_1 = 1, X_2 = 1]) = P([X_1 = 1]) \cdot P([X_2 = 1])$$

$$P([X_1 = 1, X_2 = 2]) = P([X_1 = 1]) \cdot P([X_2 = 2])$$

Meqenëse çdo probabilitet jepet si raport i çdo qelizë me N, ekuacionet e mësipërme janë llogaritur si:

$$\frac{x_{11}}{N} = \frac{x_{11}+x_{12}+x_{13}}{N} \cdot \frac{x_{11}+x_{21}}{N} \quad (2.2)$$

$$\frac{x_{12}}{N} = \frac{x_{11}+x_{12}+x_{13}}{N} \cdot \frac{x_{12}+x_{22}}{N} \quad (2.3)$$

$$\frac{x_{13}}{N} = \frac{x_{11}+x_{12}+x_{13}}{N} \cdot \frac{x_{13}+x_{23}}{N} \quad (2.4)$$

$$\frac{x_{21}}{N} = \frac{x_{21}+x_{22}+x_{23}}{N} \cdot \frac{x_{11}+x_{21}}{N} \quad (2.5)$$

$$\frac{x_{22}}{N} = \frac{x_{21}+x_{22}+x_{23}}{N} \cdot \frac{x_{12}+x_{22}}{N} \quad (2.6)$$

$$\frac{x_{23}}{N} = \frac{x_{21}+x_{22}+x_{23}}{N} \cdot \frac{x_{13}+x_{23}}{N} \quad (2.7)$$

Nga ekuacioni (2.2) dhe (2.5) marrim:

$$\frac{x_{11}}{x_{21}} = \frac{x_{11}+x_{12}+x_{13}}{x_{21}+x_{22}+x_{23}}$$

Në të njëjtën mënyrë, ekuacioni i mëposhtëm do të merret:

$$\frac{x_{11}}{x_{21}} = \frac{x_{12}}{x_{22}} = \frac{x_{13}}{x_{23}} = \frac{x_{11}+x_{12}+x_{13}}{x_{21}+x_{22}+x_{23}} \quad (2.8)$$

Pra, ka vend teorema 2.2.2:

Teorema 2.2.2. Në qoftë se dy ndryshore në tabelën e kontigjencës 2x3 janë të pavarura statistikisht, atëherë ka vend barazimi:

$$x_{11}x_{22} - x_{12}x_{21} = x_{12}x_{23} - x_{13}x_{22} = x_{13}x_{21} - x_{11}x_{23} = 0. \quad (2.9)$$

Përfundimi i arritur në tabelat e kontigjences 2x2 dhe 2x3 përgjithësohet edhe për tabelat e kontigjencës të trajtës 2xn, n> 3. Në këtë rast ekuacioni (2.9) mund të paraqitet si:

$$\frac{x_{11}}{x_{21}} = \frac{x_{12}}{x_{22}} = \frac{x_{13}}{x_{23}} = \cdots = \frac{x_{1n}}{x_{2n}} = \frac{x_{11}+x_{12}+x_{13}+\cdots+x_{1n}}{x_{21}+x_{22}+x_{23}+\cdots+x_{2n}} = \frac{\sum_{k=1}^{n} x_{1k}}{\sum_{k=1}^{n} x_{2k}} \quad (2.10)$$

Pra ka vend teorema 2.2.3:

Teorema 2.2.3. Në qoftë se dy ndryshore në tabelën e kontigjencës 2× k, k=2,3,...,n

janë të pavarura statistikisht, atëherë ka vend barazimi:

$$x_{11}x_{22} - x_{12}x_{21} = x_{21}x_{23} - x_{13}x_{22} = \cdots = x_{1n}x_{21} - x_{11}x_{n3} = 0. \qquad (2.11)$$

2.3 Krahasimi i dy raporteve

Shumë studime synojnë të krahasojnë ndryshoret me dy vlera. Le te jetë Y ndryshore me dy kategori rezultatesh që janë sukses dhe mossukses. Një tabele kontigjence 2×2 i paraqet rezultatet. Në të rreshtat janë grupime X dhe shtyllat janë kategoritë e Y.

2.3.1 Diferenca e raporteve

Për të dhënat në rreshtin i, $p_{1/i}$ është probabiliteti i suksesit në kategorinë 1. Për dy rezultate, $p_{1/i} = 1 - p_{2/i}$ dhe për thjeshtesi përdorim simbolin p_i në vend të $p_{1/i}$. *Diferenca e raporteve* (probabiliteteve) të sukseseve, $p_1 - p_2$ është krahasim bazë për dy rreshta.

Krahasimi për mossukseset është ekuivalent me krahasimin e sukseseve meqenëse

$$(1 - p_1) - (1 - p_2) = p_2 - p_1.$$

Diferenca e raporteve lëviz midis -1.0 dhe +1.0. Ajo barazohet me zero nëse rreshtat kanë shpërndarje të kushtëzuar, identike. Y është statistikisht i pavarur nga rreshtat kur $p_1 - p_2 = 0$.

Risku relativ përcaktohet (Agresti 2002) nga raporti

$$p_1/p_2 \qquad (2.12)$$

Ai duhet të jetë një numër real jonegativ. Për riskun relativ merr vlerën 1.0 i kemi pavaresinë midis X dhe Y.

Duke krahasuar rreshtat në kategorinë e dytë, marrim riskun relative si më poshtë:

$$1 - p_1/1 - p_2$$

2.3.2 Raporti i shanseve (mundësive)

Le të jetë p probabiliteti i suksesit të një ngjarjeje rasti, raporti i mundësive të suksesit me të

mossuksesin përcaktohet nga:

$$\text{odds} = p/(1-p)$$

odds është jonegative dhe odds > 1.0 kur nje sukses ka probabilitet më të madh se mossuksesi.

Nëse p= 0.75, atëhere $odds$ = 0.75⁄0.25 = 3.0; një sukses është 3 herë më i pritshem se një mossukses, ose presim për 3 suksese në çdo mossukses. Kur $odds$ =1/3, një mossukses eshte 3 herë më i pritshem se një sukses. Anasjelltas,

$$\text{p} = odds/odds + 1$$

Për shembull, kur $odds$ = 1/3, atëherë p = 0.25.

Referuar një tabele 2 x 2, me *i* rreshta, rasti i sukseseve është:

$$odds_i = p_i/1 - p_i$$

Raporti i mundësive $odds_1$ dhe $odds_2$ për 2 rreshta është

$\text{odds} = \frac{odds_1}{odds_2} = \frac{p_1(1-p_1)}{p_2(1-p_2)}$, është quajtur raport i shanseve (mundësive).

Në mënyrë të përgjithshme për *i* rreshta,

$$odds_i = p_{i1}/1 - p_{i2}\,, i = 1{,}2$$

Në këtë mënyrë raporti i mundësive jepet si më poshtë:

$$\text{odds} = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}} \qquad (2.13)$$

Nje emërtim tjetër për odds është raporti i prodhimit të tërthortë.

2.3.3 Lidhja midis Raportit të mundësive dhe Riskut Relativ

Nga barazimet e mesipërme (2.12) dhe (2.13) nxjerrim se

$$raporti\ i\ mundësive\ =\ risku\ relativ\ \left(\frac{1-p_2}{1-p_1}\right).$$

Madhesitë janë të krahasueshme nëse probabiliteti p_i i rezultateve është afër 0 për të dyja grupet.

2.3.4 Pavarësia statistikore e tabelave të kontigjencës mxn

Konsiderojmë tabelën e kontigjencës mxn, si në tabelen 2.3 Atëherë pavarësia statistikore midis X_1, X_2 jepet me formulën:

$\mathrm{P}\left(\left[X_1 = A_i, X_2 = B_j\right]\right) = P([X_1 = A_i])P\left(\left[X_2 = B_j\right]\right)$

$((i = 1,2, \dots, m \; dhe \; j = 1,2, \dots, n)$

Meqenëse cdo probabilitet jepet si raport i cdo qelize me N atëherë ekuacionet e mësipërme llogariten si:

$$\frac{x_{ij}}{N} = \frac{\sum_{i=1}^{n} x_{ik}}{N} \frac{\sum_{j=1}^{m} x_{ik}}{N}$$

nga ku marim:

$$x_{ij} = \frac{\sum_{i=1}^{n} x_{ik} \sum_{j=1}^{m} x_{lj}}{N}$$

Për një j të fiksuar kemi:

$$\frac{x_{i_a j}}{x_{i_b j}} = \frac{\sum_{k=1}^{n} x_{i_a k}}{\sum_{k=1}^{n} x_{i_b k}}$$

Për një i të fiksuar kemi:

$$\frac{x_{i j_a}}{x_{i j_b}} = \frac{\sum_{l=1}^{m} x_{l j_a}}{\sum_{l=1}^{m} x_{l j_b}}$$

Ky ekuacion ka vend për cdo j, prandaj merret ekuacioni i mëposhtëm:

$$\frac{x_{i_a 1}}{x_{i_b 1}} = \frac{x_{i_a 2}}{x_{i_b 2}} = \cdots = \frac{x_{i_a n}}{x_{i_b n}} = \frac{\sum_{k=1}^{n} x_{i_a k}}{\sum_{k=1}^{n} x_{i_b k}}$$

Vemë re se ana e djathtë e barazimit është madhësi konstante atëherë dhe ana e majtë është e tillë. Në këtë kushte ka vend teorema:

Teorema 2.3.1. Nëse dy ndryshore në tabelën e kontigjencës të trajtës mxn, (2.3) janë të pavarura statistikisht, atëherë ka vend barazimi:

$$\frac{x_{i_a 1}}{x_{i_b 1}} = \frac{x_{i_a 2}}{x_{i_b 2}} = \cdots = \frac{x_{i_a n}}{x_{i_b n}} = konstante, (i_a, i_b = 1,2, \dots, m)$$

2.4 Matricat e kontigjencës

Fillimisht përkufizohet matrica e kontigjencës dhe rangu i saj.

Përkufizim 2.4.1. Matrica e kontigjencës $M_{X_{A,B}}$ është matrica e cila ka si elementë, elementë të tabelës së kontigjencës $X_{A,B}$ të dy ndryshoreve A dhe B duke përjashtuar vlerat margjinale.

Përkufizim 2.4.2. Rangu i tabelës së kontigjencës është rangu i matricës së kontigjencës.

Matrica e kontigjencës e tabelës së kontigjencës $m \times n$ është (X(X_1, X_1)) e përcaktuar si $M_{X_{1,X_2}}$ si më poshtë:

$$M = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}.$$

2.4.1 Pavarësia e tabelës së kontigjencës 2x2

Duke iu referuar tabelës së kontigjencës (2x2) matrica korresponduese e kontigjencës $M_{X_{1,X_2}}$ është: $M = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$.

Vihet re se përcaktori i matricës së kontigjencës është:

$\det(M_{X_{X_{1,X_2}}}) = x_{11}x_{22} - x_{12}x_{21}$

Rangu i matricës së kontigjencës 2x2 është: rang = $\begin{cases} 2, nëse\ \det(M_{X_{1,X_2}}) \neq 0 \\ 1, nëse\ \det(M_{X_{1,X_2}}) = 0 \end{cases}$

Nga teorema 2.2.1 marret teorema 2.4.1:

Teorema 2.4.1. Nëse rangu i matricës korresponduese të tabelës së kontigjencës 2x2 është i barabartë me 1, arëherë dy ndryshoret janë të pavarura statistikisht.

rang = $\begin{cases} 1, të\ pavarura\ statistikisht \\ 2, të\ varura \end{cases}$

Ky arsyetim përgjithësohet dhe në rastin e tabelës së kontigjencës $2 \times n$

Teorema 2.4.2. Nëse rangu i matricës korresponduese të tabelës së kontigjencës 2xn është i barabartë me 1, arëherë dy ndryshoret janë të pavarura statistikisht.

rang = $\begin{cases} 1, \text{të } pavarura\ statistikisht \\ 2, \text{të } varura \end{cases}$

2.4.2 Pavarësia e tabelës së kontigjencës 3x3

Në tabelat e kontigjencës kur numri i rrjeshtave dhe shtyllave është më i madh se 3 situata ndryshon. Vihet re se kur rangu i matricës korresponduese është 1, ndryshoret janë të pavarura statistikisht dhe në rastin kur rangu është $min(m,n)$, atëherë ndryshoret janë të varura. Por shtrohet problemi se cfarë ndodh kur rangu i matricës eshtë më i madh se 1 dhe më i vogël se $min(m,n)$? Në këto kushte merret matrica e kontigjencës të tabelës së kontigjencës 3x3.

Shembull 3.2. Supozojmë se matrica korresponduese është:

$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$, përcaktori i matricës A, $\det(A) = 1 \cdot (-1)^{1+1} \cdot det\begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix} + 2 \cdot (-1)^{1+2} \cdot$

$$det\begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix} + 3 \cdot (-1)^{1+3} \cdot det\begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} = 1 \cdot (-3) + 2 \cdot 6 = 3 \cdot (-3) = 0.$$

Pra, rangu i matricës A nuk është 3, dmth ai është më i vogël se 3. Nga ana tjetër meqë (123) ≠k(456) dhe (123) ≠k(789), atëherë rangu i matricës nuk është 1. Pra rangu i matricës A është 2. Në këtë rast kemi varësi lineare midis rrjeshtave domethënë se njëri nga rrjeshtat shprehet si kombinim linearë i dy rrjeshtave të tjerë.

Për shembull, $(456) = \frac{1}{2}\{(123) + (789)\}$

Prandaj, në këtë rast, mund të themi se dy nga tre ciftet e një ndryshoreje janë të varur nga ndryshorja tjetër, por një cift është statistikisht i pavarur nga cifti tjetër në lidhje me kombinim linear dy çifte të tjera. Është e lehtë për të parë se ky rast përfshin rastet kur dy cifte janë statistikisht të pavarur nga ndryshorja tjetër, por tabela bëhet statistikisht e varur me ndryshoren tjetër. Pra, matrica përkatëse është një përzierje e varësisë statistikore dhe pavarësisë. Këtë rast e quajmë pavarësi kontekstuale. Në këtë rast kemi të bëjmë me pavarësi kontekstuale.

Teorema 2.4.3. Në qoftë se rangu i matricës korresponduese të tabelës së kontigjencës 3x3 është 1, atëherë dy ndryshoret e tabelës së kontigjencës janë të pavarura statistikisht.

Pra, $rang = \begin{cases} 1, & \text{të } pavarura\ statistikisht \\ 2, & kontekstualisht\ \text{të } varur \\ 3, & \text{të } varur \end{cases}$.

Kjo situate mund të përgjithësohet dhe në rastin e tabelës së kontigjencës mxn.

2.4.3 Shpërndarjet margjinale dhe matrica e vlerave të pritshme

Në analizat statistikore të një tabele kontigjence, shpërndarja margjinale luan një rol të rëndësishëm. Sidomos, shpërndarja margjinale është e lidhur ngushtë me pavarësinë statistikore midis dy ndryshoreve. Në fillim përkufizohet shpërndarja margjinale rrjesht ose shtyllë.

Përkufizim 2.4.1. Shpërndarje margjinale rrjesht të $M_{X_{X_1,X_2}}$, është mr_{M,X_2} e cila përcaktohet si:

$$x_{.j} = (x_{.1}, x_{.2}, \dots, x_{.m}),$$

ku është marrë si prodhim i

$$(1\ 1 \qquad 1) \cdot \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

Shpërndarje margjinale shtyllë të $M_{X_{X_1,X_2}}$,është mc_{M,X_1} e cila përcaktohet si:

$$x_{i.} = \begin{pmatrix} x_{1.} \\ \vdots \\ \vdots \\ x_{n.} \end{pmatrix}$$

ku është marrë si prodhim i:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}$$

Është e dukshme se shpërndarja margjinale rrjesht dhe shtyllë korrespondon me

$P([X_1 = A_i])$ dhe $P\big([X_2 = B_j]\big)$.

Atëherë, kur kemi pavarësi statistikore midis dy ndryshoreve kemi:

$\mathrm{P}\big([X_1 = A_i, X_2 = B_j]\big) = P([X_1 = A_i]) \cdot P\big([X_2 = B_j]\big)$

Ku cdo vlerë x_{ij} është marrë si shumëzim i $x_{i.}$ dhe $x_{.j}$

$$\frac{x_{ij}}{x_{..}} = \frac{x_{i.}}{x_{..}} \cdot \frac{x_{.j}}{x_{..}}$$

Kështu matrica e vlerave të pritshme është matrica elementët e së cilës janë dhënë nga pavarësia statistikore e dy ndryshoreve e cila është përcaktuar si më poshtë:

Përkufizim 2.4.2. Matrica e vlerave të pritjes $E_{X_1,X_2}(m,n,N)$ përcaktohet nga $M_{X_1,X_2}(m,n,N)$.

Matrica përcaktohet me anë të elementëve e_{ij} që e_{ij} jepen si prodhim i elementëve të shpërndarjes margjinale

$$e_{ij} = \frac{x_{i.} \cdot x_{.j}}{x_{..}}$$

Pra, është e qartë se kjo matricë mund të zbërthehet si shumëzim i shpërndarjeve margjinale rrjesht dhe shtyllë.

Teorema 2.4.4. Matricae pritjes $E_{X_1,X_2}(m,n,N)$ zbërthehet si:

$$E_{X_1,X_2}(m,n,N) = mr_{M,X_2} \times mc_{M,X_1} \qquad (2.14)$$

Testin Hi-katror paraqitet si shumë e differences midis vlerës së vrojtuar dhe vlerës së pritur:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{(x_{ij} - \frac{x_{i.} \cdot x_{.j}}{x_{..}})^2}{\frac{x_{i.} \cdot x_{.j}}{x_{..}}}$$

e cila ka shpërndarje χ^2 me $(m-1)(n-1)$ shkallë lirie.

Shënojmë diferencën midis vlerës së vrojtuar dhe vlerës së pritur me σ_{ij}. Nga pikëpamja e teorisë së matricave σ_{ij} është përcaktuar nga

$$M_{X_1,X_2}(m,n,N) - E_{X_1,X_2}(m,n,N)$$

ku kjo matricë quhet matricë e mbetjes.

Përkufizim 2.4.3. Matrica e mbetjes $S_{X_1,X_2}(m,n,N)$ është përcaktuar si:

$$(\sigma_{ij})$$

ku σij merret si një diferencë e elementëve x_{ij} me elementët e_{ij} .

$$(\sigma_{ij}) = x_{ij} - \frac{x_{i.} \cdot x_{.j}}{x_{..}} \qquad (2.15)$$

2.5 Shpërbërja e matricave të mbetjes

Matrica e mbetjes zotëron karakteristika shumë të rëndësishme.

Teorema 2.5.1. Le të jetë $S_{X_1,X_2}(m,n,N)$ matricë e mbetjes së matricës $M_{X_1,X_2}(m,n,N)$. Atëherë $\det(S_{X_1,X_2}(m,n,N)) = 0$.

Vërtetim

Nga ekuacioni (2) për cdo ($j = 1,2,\dots,n$),

$$\sigma_{.j} = \sum_{i=1}^{n} \sigma_{ij} = \sum_{i=1}^{n} (x_{ij} - \frac{x_{i.} \cdot x_{.j}}{x_{..}}) = x_{.j} - \frac{x_{..} \cdot x_{.j}}{x_{..}} = 0.$$

Kështu, $\sigma_{1j} + \sigma_{2j} + \cdots + \sigma_{mj} = 0$,

Ku të paktën një rresht mund të shprehet si kombinimi linearë i rrjeshtave të tjerë pra përcaktori i kësaj matrice është e barabartë me zero.

Teorema 2.5.1 tregon se rangu matricës $S_{X_1,X_2}(m,n,N)$ është të shumtën $\min(m,n) - 1$.

Në këto kushtë kemi një pavarësi kontekstuale tek tabela e kontigjencës. Kur matrica e kontigjencës jep pavarësi kontekstuale atëherë matrica mund të zbërthehet si prodhim i tre matricave:

$$M(m,n,N) = M_{m\times s} M(s,s,N') M_{s\times n}$$

Ku $M(s,s,N')$ është matricë bazë e $M(m,n,N)$ rangu i së cilës është i barabartë me s $M_{m\times n}$ është një matricë me përmasë $m \times n$.

Për më tepër matrica e pritjes mund të përgjithësohet si tek ekuacioni 2.14 në këto kushte merret një shpërbërje e matricës së kontigjencës. Këtë ka vend teorema:

Teorema 2.5.2. $M_{X_1,X_2}(m,n,N) = mr_{M,X_2} \times mc_{M,X_1} + M_{m\times s} M'(s,s,N') M_{s\times n}$

Ku $M'(s,s,N')$ është matrica bazë e $S_{X_1,X_2}(m,n,N)$, ($s \le \min(m,n)$).

Shembull 5.1.

Le të konsiderojmë matricën e mëposhtme

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Atëherë matrica e pritjess E_A është: $E_A = \frac{1}{3}\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix}$

Matrica e mbetjes është: $S_A = \frac{1}{3}\begin{pmatrix}2 & -1 & -1\\ -1 & 2 & -1\\ -1 & -1 & 2\end{pmatrix}$ ku det(S_A) = 0.

Vemë re se rangu i matricës është 2.

Kështu matrica e vlerave të pritshme dhe matrica e mbetjeve mund të shpërbëhen si më poshtë:

$$E_A = \frac{1}{3}\begin{pmatrix}1\\ 1\\ 1\end{pmatrix}(1 \quad 1 \quad 1)$$

dhe

$$S_A = \frac{1}{3}\begin{pmatrix}1 & 0\\ 0 & 1\\ -1 & -1\end{pmatrix}\begin{pmatrix}2 & -1\\ -1 & 2\end{pmatrix}\begin{pmatrix}1 & 0 & -1\\ 0 & 1 & -1\end{pmatrix}.$$

2.5.1 Shpërbërja e matricave 2×2

Një nga rezultatet e Teoremës 2.5.2 është shpërbërja e matricës 2×2, para se ta bëjmë këtë shpërbërje le të japim rastin e përgjithshëm.

Teorema 2.5.3. Mbetje e $M_{X_1,X_2}(m, n, N)$ merret

$$\sigma_{ij} = \frac{1}{x_{..}}\{x_{ij}\sum_{k\neq i}\sum_{l\neq j}x_{kl} - (\sum_{l\neq j}x_{il})(\sum_{k\neq i}x_{kj})\}.$$

Vërtetim.

$$\sigma_{ij} = x_{ij} - \frac{x_{i.}\cdot x_{.j}}{x_{..}} = \frac{1}{x_{..}}\left\{x_{ij}(x_{.j} + \sum_{l\neq j}x_{.l}) - (x_{ij} + \sum_{l\neq j}x_{il})x_{.j}\right\}$$

$$= \frac{1}{x_{..}}\left\{x_{ij}(x_{.j} + \sum_{k\neq i,l\neq j}x_{kl} + \sum_{l\neq j}x_{il}) - (\sum_{l\neq j}x_{il})(x_{ij} + \sum_{k\neq i}x_{kj})\right\}$$

$= \frac{1}{x_{..}} \{ x_{ij} \sum_{k \neq i} \sum_{l \neq j} x_{kl} - (\sum_{l \neq j} x_{il}) (\sum_{k \neq i} x_{kj}) \}.$

Nga kjo teoremë kemi këto rezultate për matricat 2×2.

Teorema 2.5.4. $S(2,2,N)$ tregon mbetje $M(2,2,N)$.

Atëherë

$$\sigma_{ij} = \frac{(-1)^i (-1)^j}{x_{..}} (x_{12}x_{21} - x_{11}x_{22}) = \frac{(-1)^{i+j}}{x_{..}} (x_{12}x_{21} - x_{11}x_{22}).$$

Vërtetim.

Nga teorema 2.5.2, marrim: $i = j = 1$

dhe $\sigma_{11} = \frac{1}{x_{..}} \{ x_{11} \sum_{k \neq 1} \sum_{l \neq 1} x_{kl} - (\sum_{l \neq 1} x_{1l}) (\sum_{k \neq 1} x_{k1}) \} =$

$$= \frac{1}{x_{..}} (x_{11}x_{22} - x_{12}x_{21}).$$

Në të njetën mënyrë:

$$\sigma_{12} = \frac{1}{x_{..}} (x_{12}x_{21} - x_{11}x_{22})$$

$$\sigma_{21} = \frac{1}{x_{..}} (x_{12}x_{21} - x_{11}x_{22})$$

$\sigma_{22} = \frac{1}{x_{..}} (x_{11}x_{22} - x_{12}x_{21})$ në këto kushte matrica mbetje jepet në këtë mënyrë:

$$\frac{x_{11}x_{22} - x_{12}x_{21}}{x_{..}} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Pra, në rastin e matricës 2×2 ka vend kjo teorema 2.5.5

Teorema 2.5.5. Shpërbërja e $M(2,2,N)$ jepet si:

$$M(2,2,N) = E_M(2,2,N) + S_M(2,2,N)$$

$$= \frac{1}{x_{..}} \begin{pmatrix} x_{1.} \\ x_{2.} \end{pmatrix} (x_{.1} \quad x_{.2}) + \frac{x_{11}x_{22} - x_{12}x_{21}}{x_{..}} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Shembull 5.2.

Le të marrim matricën $B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

Matrica e vlerave të pritjes dhe matrica e mbetjes janë:

$$E_B = \frac{1}{10}\begin{pmatrix} 12 & 18 \\ 28 & 42 \end{pmatrix}, S_B = \frac{1}{10}\begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}$$

Kështu këto dy matrica E_B, S_B shpërbëhen si më poshtë:

$$E_B = \frac{1}{10}\begin{pmatrix} 3 \\ 7 \end{pmatrix}(4 \quad 6) \text{ dhe } S_B = \frac{2}{10}\begin{pmatrix} -1 \\ 1 \end{pmatrix}(-1 \quad 1)$$

Në këtë mënyrë marica e dhënë B shpërbëhet në këtë mënyrë:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \frac{1}{10}\begin{pmatrix} 3 \\ 7 \end{pmatrix}(4 \quad 6) + \frac{2}{10}\begin{pmatrix} -1 \\ 1 \end{pmatrix}(-1 \quad 1).$$

2.6 Shpërbërja e matricave 3×3

2.6.1 Matrica 3×3 me rang 2

Meqenëse jemi në matrica me përmasa $\mathbf{3 \times 3}$ dhe rangu i të cilës është 2 atëherë për konkretizim, supozohet se rrjeshti i tretë shprehet si kombinim linearë i dy rrjeshtave të tjerë ose shtylla e parë shprehet si kombinim linearë i dy shtyllave të tjera pra në këto kushte kemi:

$$(x_{31}\, x_{32}\, x_{33}) = s(x_{11}\, x_{12}\, x_{13}) + (x_{21}\, x_{22}\, x_{23})$$

dhe
$$\begin{pmatrix} x_{13} \\ x_{23} \\ x_{33} \end{pmatrix} = u\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix} + v\begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \end{pmatrix}.$$

Vemë re se rangu i matricës së mbetjes është të shumtën 2, atëherë këto katër elementë janë mjaftueshëm për të studiuar karakteristikat e matricës së mbetjes.

Bazuar në teoremën 2.4.4 njehsojmë σ_{11}

$$\sigma_{11} = \frac{1}{x_{..}}\left\{x_{ij}\sum_{k\neq 1}\sum_{l\neq 1} x_{kl} - \left(\sum_{l\neq 1} x_{1l}\right)\left(\sum_{k\neq 1} x_{k1}\right)\right\}$$
$$= \frac{1}{x_{..}}\{x_{11}(x_{22} + x_{23} + x_{32} + x_{33}) - (x_{11} + x_{12})(x_{21} + x_{31})\}$$

$$= \frac{1}{x_{..}}\{x_{11}(x_{22} + x_{23}) + (sx_{12} + tx_{22}) + (sx_{13} + tx_{23})) - (x_{12} + x_{13})(x_{21} + sx_{11} + tx_{21})\}$$

$$= \frac{1}{x_{..}}(1+t)\{x_{11}x_{22} + x_{11}x_{23} - (x_{12} + x_{13})x_{21}\}$$

$$= \frac{1}{x_{..}}(1+t)\{x_{11}x_{22} + x_{11}(ux_{21} + vx_{22}) - x_{21}(x_{12} + ux_{11} + vx_{12})\}$$

$$= \frac{1}{x_{..}}(1+t)(1+v)(x_{11}x_{22} - x_{12}x_{21}).$$

Në mënyrë të njetë marrim:

$$\sigma_{12} = \frac{1}{x_{..}}(1+t)(1+u)(x_{12}x_{21} - x_{11}x_{22})$$

$$\sigma_{21} = \frac{1}{x_{..}}(1+s)(1+v)(x_{12}x_{21} - x_{11}x_{22})$$

$$\sigma_{22} = \frac{1}{x_{..}}(1+s)(1+u)(x_{11}x_{22} - x_{12}x_{21})$$

Atëherë matrica e mbetur paraqitet si më poshtë:

$$\frac{1}{x_{..}}(x_{11}x_{22} - x_{12}x_{21}) \cdot \begin{pmatrix} (1+t)(1+v) & -(1+t)(1+u) \\ -(1+s)(1+v) & (1+s)(1+u) \end{pmatrix}$$

Kështu rangu i matricës së mbetjes është i barabartë me 1.

Ka vend teorema 2.6.1:

Teorema 2.6.1. Në qoftë se rangu i matricës së kontigjencës 3x3 është 2, atëherë rangu i matricës së mbetjes është 1.

Shembull 6.1.

Jepet matrica C, $C = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$.

det(C) = 0, vemë re se rangu i C është 2.

Gjejmë matricën e vlerave të pritjes dhe matricën e mbetjes, të cilat janë:

$E_A = \frac{1}{45}\begin{pmatrix} 72 & 90 & 108 \\ 180 & 225 & 270 \\ 288 & 360 & 432 \end{pmatrix}$ dhe $S_A = \frac{1}{45}\begin{pmatrix} -27 & 0 & 27 \\ 0 & 0 & 0 \\ 27 & 0 & -27 \end{pmatrix}$, ku rangu i S_A është i barabartë me

1, (rg(S_A) = 1).

Kështu matrica e vlerave të pritjes dhe matrica e mbetjes shpërbëhen në këtë mënyrë:

$$E_A = \frac{1}{45}\begin{pmatrix}6\\15\\24\end{pmatrix}(12 \quad 15 \quad 18)$$

dhe

$$S_A = \frac{1}{45}\begin{pmatrix}-1\\0\\1\end{pmatrix}(-1 \quad 0 \quad 1)$$

2.6.2 Matrica 3 × 3 me rang 3

Në rastin kur kemi këtë matricë dhe rangu i së cilës është 3, atëherë rangu i matricës së mbetjes është 2. Le ta konkretizojmë këtë situatë.

Shembull 6.2

Jepet matrica D, $D = \begin{pmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 10\end{pmatrix}$.

Matrica e vlerave të pritjes dhe matrica e mbetjes gjenden në këtë mënyrë:

$$E_A = \frac{1}{46}\begin{pmatrix}72 & 90 & 114\\180 & 225 & 285\\300 & 375 & 475\end{pmatrix}$$

dhe

$$S_A = \frac{1}{46}\begin{pmatrix}-26 & 2 & 24\\4 & 5 & -9\\22 & -7 & 15\end{pmatrix}$$

ku rangu i S_A është i barabartë me 2, (rg(S_A) = 2).

Kështu këto matrica mund të shpërbëhen si më poshtë:

$$E_A = \frac{1}{46}\begin{pmatrix}6\\15\\25\end{pmatrix}(12 \quad 15 \quad 19)$$

dhe

$$S_A = \frac{1}{46}\begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 5 & -9 \\ -7 & -15 \end{pmatrix}$$

2.7 Tabelat e kontigjencës me tre hyrje

Le të shënojmë me ■ shumën e rrjeshtave ose të shtyllave të matricës së kontigjencës. Kjo jepet sipas Tsumoto (2006)

$$x_{i\blacksquare} = \sum_{j=1}^{n} x_{ij} \qquad (2.16)$$

$$x_{j\blacksquare} = \sum_{i=1}^{m} x_{ij} \qquad (2.17)$$

Ku (2.16) dhe (2.17) tregojnë shumat e vlerave margjinale rrjesht dhe shtyllë ku vemë re se

$x_{\blacksquare\blacksquare} = N$

Ku, N është vëllimi total i zgjedhjes.

Atëherë ekuacioni $x_{ij} = \frac{\sum_{i=1}^{n} x_{ik} \sum_{j=1}^{m} x_{lj}}{N}$ mund të paraqitet si:

$$\frac{x_{ij}}{x_{\blacksquare\blacksquare}} = \frac{x_{i\blacksquare}}{x_{\blacksquare\blacksquare}} \cdot \frac{x_{\blacksquare j}}{x_{\blacksquare\blacksquare}} \qquad (2.18)$$

Ku merret $x_{ij} = \frac{x_{i\blacksquare} \cdot x_{\blacksquare j}}{x_{\blacksquare\blacksquare}}$

ose $x_{ij} x_{\blacksquare\blacksquare} = x_{i\blacksquare} x_{\blacksquare j}$

Kështu, pavarësia statistikore shikohet si specifike për marrëdhëniet ndërmjet i, j dhe "■". Me përdorimin lidhjes së mësipërme , ekuacionit (1) mund të përshkruhet dhe si:

$$\frac{x_{i_1 j}}{x_{i_2 j}} = \frac{x_{i_1\blacksquare}}{x_{i_2\blacksquare}}$$

ku ana e djathtë jep raportin e shumës së shtyllës margjinale. Ekuacioni (2.18) mund të zgjerohet në raste multinomial, le të marrim parasysh një rast me tre hyrje kështu, Pavarësia Statistikore me tre atributet është përcaktuar si:

$$\frac{x_{ijk}}{x_{\blacksquare\blacksquare\blacksquare}} = \frac{x_{i\blacksquare\blacksquare}}{x_{\blacksquare\blacksquare\blacksquare}} \cdot \frac{x_{\blacksquare j\blacksquare}}{x_{\blacksquare\blacksquare\blacksquare}} \frac{x_{\blacksquare\blacksquare k}}{x_{\blacksquare\blacksquare\blacksquare}} \qquad (2.19)$$

Ku marrrim

$$x_{ijk}x_{\blacksquare\blacksquare\blacksquare}{}^{2} = x_{i\blacksquare\blacksquare}x_{\blacksquare j\blacksquare}x_{\blacksquare\blacksquare k} \qquad (2.20)$$

E cila korrespondon me:

$P(A = a, B = b, C = c) = P(A = a)P(B = b)P(C = c)$ (2.21)

ku A, B, C i korrespondojnë emrave të hyrjeve përkatësisht i, j, k.

Pavarësia statistikore në tabelat e kontigjencës me tre hyrjeve kërkon që të gjitha tabelat me dy hyrje nga tre hyrje duhet të plotësojë kushtin e pavarësisë statistikore. Kështu, për ekuacioni (2.21), ekuacionet e mëposhtme duhet të plotësojnë:

$$P(A = a, B = b) = P(A = a)P(B = b)$$

$$P(B = b, C = c) = P(B = b)P(C = c)$$

$$P(A = a, C = c) = P(A = a)P(C = c)$$

Kështu

$$x_{ij\blacksquare}x_{\blacksquare\blacksquare\blacksquare} = x_{i\blacksquare\blacksquare}x_{\blacksquare j\blacksquare} \qquad (2.22)$$

$$x_{i\blacksquare k}x_{\blacksquare\blacksquare\blacksquare} = x_{i\blacksquare\blacksquare}x_{\blacksquare\blacksquare k} \qquad (2.23)$$

$$x_{\blacksquare jk}x_{\blacksquare\blacksquare\blacksquare} = x_{\blacksquare j\blacksquare}x_{\blacksquare\blacksquare k} \qquad (2.24)$$

Nga ekuacioni (2.20) dhe (2.22) marrim

$$x_{ijk}x_{\blacksquare\blacksquare\blacksquare} = x_{ij\blacksquare}x_{\blacksquare\blacksquare k}$$

Nga ku,

$$\frac{x_{ijk}}{x_{ij\blacksquare}} = \frac{x_{\blacksquare\blacksquare k}}{x_{\blacksquare\blacksquare\blacksquare}} \qquad (2.25)$$

$$\frac{x_{ijk}}{x_{i\blacksquare k}} = \frac{x_{\blacksquare j\blacksquare}}{x_{\blacksquare\blacksquare\blacksquare}} \qquad (2.26)$$

$$\frac{x_{ijk}}{x_{\blacksquare jk}} = \frac{x_{i\blacksquare\blacksquare}}{x_{\blacksquare\blacksquare\blacksquare}} \qquad (2.27)$$

Përfundimet e mësipërme në trajtë të përmbledhur paraqiten ne teoremën 2.7.1

Teorema 2.7.1: Në qoftë se nje tabelë kontigjence me tre hyrje kënaq kushtin e pavarësisë statistikore atëherë kanë vend këto barazime

$$\frac{x_{ijk}}{x_{ij\blacksquare}} = \frac{x_{\blacksquare\blacksquare k}}{x_{\blacksquare\blacksquare\blacksquare}}$$

$$\frac{x_{ijk}}{x_{i\blacksquare k}} = \frac{x_{\blacksquare j\blacksquare}}{x_{\blacksquare\blacksquare\blacksquare}}$$

$$\frac{x_{ijk}}{x_{\blacksquare jk}} = \frac{x_{i\blacksquare\blacksquare}}{x_{\blacksquare\blacksquare\blacksquare}}$$

Konkluzione:

Nëse tre ndryshore në tabelën e kontigjencës të trajtës mxn, janë të pavarura statistikisht, atëherë kanë vend barazimet:

$$\frac{x_{ijk_a}}{x_{ijk_b}} = \frac{x_{\blacksquare\blacksquare k_a}}{x_{\blacksquare\blacksquare k_b}}$$

$$\frac{x_{ij_ak}}{x_{ij_bk}} = \frac{x_{\blacksquare j_a\blacksquare}}{x_{\blacksquare j_b\blacksquare}}$$

$\frac{x_{i_ajk}}{x_{\blacksquare i_bjk}} = \frac{x_{i_a\blacksquare\blacksquare}}{x_{i_b\blacksquare\blacksquare}}$ për cdo i, j, k.

2.8 Vlerësimi i qelizave probabilitare në tabelat e kontigjencës

2.8.1 Vlerësimi i parametrave binomial

Për r modele të pavarura me shpërndarje binomiale, tabela e kontigjencës ka $r \times 2$ përmasë. Për thjeshtësi, shënojmë $\{\pi_i\}$ parametrat binomial.

Ndodh shpesh që vlerësimi i parametrave binomial kryhet duke përdorur vlerësimin empirik Baysian. Sipas Griffin (1971) supozohet për një prior të panjohur në parametrat e shpërndarjes Binomiale. Vlerësimi Bayes-ian shprehet në një formë që nuk përfshin priorin në mënyrë eksplicite por ai në termat e probabiliteteve të ngjarjeve përfshin provat binomiale. Një vlerësim altenativë është përdorimi i një vlerësimi hiearkik. Leonard (1972) supozon se $\{logit(\pi_i)\}$ janë të pavarura nga shpërndarja normale $N(\mu, \sigma^2)$, gjithashtu ai supozon një prior uniform për μ dhe shpërndarje prior hi-katror për σ^2.

Veçanërisht, ai supozon se $\nu\lambda/\sigma^2$ është e pavarur nga μ dhe ka shpërndarje hi-katror (χ^2) me shkallë lirie $df = \nu$, ku λ është një vlerësues prior i σ^2 dhe ν është masa e sigurisë së priorit. Për thjeshtësi, ai përdori një prior uniform të përshtatshëm për $\log(\sigma^2)$. Duke integruar në lidhje me μ dhe σ^2, vlerësimi i tij korespondon me një prior multivariant për $\{\text{logit}(\pi_i)\}$. Për zgjedhjen propocional $\{p_i\}$ dhe mesataren e peshave të $\{logit(p_j)\}$.

Berry dhe Christensen (1976) përdorën shpërndarjen prior të $\{\pi_i\}$ për të përftuar një proçes prior Dirichle. Një mënyrë e kësaj proçedure

a) me $r = 2$ është masa në shumën e katrorve e cila është mesatarja e peshave e produktit të dy shpërndarjeve beta, ku një shpërndarje beta është përqëndruar në atë pozicion ku $\pi_1 = \pi_2$. Posteriori është një proçes i miksuar Dirichle.

b) Në rastin kur $r > 2$ ose $r > 3$, njehsimi bëhet kompleks.

Albert dhe Gupa (1983) përdorën një vlerësim hiearkik me prior të pavarur beta $(\alpha, k - \alpha)$ në parametrat binomial $\{\pi_i\}$ ku çdo prior i hapit të dytë ka shpërndarje uniforme diskrete

$$\pi(\alpha) = 1/(k-1), \alpha = 1, \dots, k-1.$$

Në rezultatin e priorit kufitar të $\{\pi_i\}$, përmasa K përcakton shtrirjen e korrelacionit ndërmjet $\{\pi_i\}$.

2.8.2 Vlerësimi i qelizave probabilitare multinomiale

Le të konsiderojmë tabelat e kontigjencës me përmasë $r \times c$, me qeliza të numërueshme $\{n_{ij}\}$ dhe probabilitete $\{\pi_{ij}\} = \pi$. Më tej, kjo ide përdoret në tabelat e kontigjencës me cdo dimension. Nga Finberg dhe Holand (1972) propozohen vlerësues të $\{\pi_{ij}\}$ duke përdorur prior me të dhëna të pavarura. Për një zgjedhje të vecantë Dirichle $\{\gamma_{ij}\}$

Për vlerësimin Bays-ian

$$[n/(n+k)]p_{ij} + [k/(n+k)]\gamma_{ij}$$

Tregohet se minimumi i gabimit mesatar katror merret për

$$k = \left[1 - \sum {\pi_{ij}}^2\right] / \sum \left(\gamma_{ij} - \pi_{ij}\right)^2.$$

Vlerat optimale të $k = k(\gamma, \pi)$ varen nga π dhe përdorin vlerësimin $k(\gamma, p)$. Kur p ndodhet afër priorit të supozuar γ, $k(\gamma, p)$ rritet dhe priori i supozuar merr më shumë peshë në vlerësimin posterior. Më tej, selektohen $\{\gamma_{ij}\}$ bazuar në përshtatshmërinë e modelit të thjeshtuar.

Epstein dhe Fienberg (1992) sugjerojnë ndjekjen e dy hapave në qelizat probabilitare.

1. Vendosjen e një priori Dirichle $k(\gamma, p)$ në π dhe përdorimin e një parametrizimi loglinear për $\{\gamma_{ij}\}$.

2. Vendosjen e një shpërndarje prior normal multivariant në termat e modelit loglinear të priorit $\{\gamma_{ij}\}$.

Zbatimi i parametrizimit loglinear të priorit $\{\gamma_{ij}\}$ tek qelizat probabilitare $\{\pi_{ij}\}$ pranon dhe pasiguri rreth strukturës loglineare për $\{\pi_{ij}\}$.

Albert (1987) sugjeron vlerësimet Bays-iane empirike duke përdorur priorët e miksuar. Për qelizat e numërueshme $n = \{n_{ij}\}$, përftohen vlerësues posterior si më poshtë:

$$E(\pi_{ij}/\boldsymbol{n}, k) \approx n_{ij} + k\widetilde{\pi_{ij}}/(n+k), \qquad \widetilde{\pi_{ij}} = p_i + p_j$$

Ku vlerësimi i K -së të bëhet nga densiteti i kushtëzuar $m\ (n/k)$.

2.8.3 Vlerësimi i parametrave të modelit loglinear në tabelat me dy hyrje

Le të konsiderojmë një tabelë kontigjence me përmasë $r \times c$. Për probabilitetin multinomial Lindley (1964) përdori një shpërndarje prior Drichle. Ai tregon se dallimi i qelizave log probabilitare të tilla si klasa log probabilitare, kanë vlerësim të përbashkët shpërndarjen normale si shpërndarje posterior (për vëllim të mëdha të zgjedhjes), e cila jep vlerësim Bayes-ian analog (ose të njetë) me rezultatet e dendurisë standarte të tabelave të kontigjencës me dy hyrje.

Block dhe Watson (1967) duke përdorur të njetën strukturë të Lindley (1964) përmisuan vlerësuesit e shpërndarjes posterior dhe përdorën kombinimet lineare në qelizat probabilitare. Sic e kemi përmendur më lart. Një disavantazh i një priori Dirichle është se ai nuk lejon vendosjen e strukturës probabilitare si në modelin loglinear.

Leonard (1975) për modelin loglinear fokusohet në parametrat e modelit

$$log[E(n_{ij})] = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

Ku prior është me shpërndarje normale. Ai tregoi se këmbyeshmëria brenda cdo bashkësie të parametrave loglinear gjen më tepër përdorim se këmbyeshmëria e probabiliteteve multinomial që përftohet me një prior Dirichle. Duke supozuar rezultatet e rreshtit $\{\lambda_i^X\}$ rezultatet e kolonës $\{\lambda_i^Y\}$ dhe rezultatet e bashkëveprimit $\{\lambda_{ij}^{XY}\}$ janë prior të pavarur. Për secilën nga këto tre bashkësi duke i dhënë mesatares vlerën μ dhe dispersion σ^2 përftohet:

1) Prior i pavarur dhe me shpërndarje normale $N(\mu, \sigma^2)$.

2) μ supozohet me shpërndarje uniforme dhe σ^2 supozohet me shpërndarjen inverse χ^2.

Për lehtësi veprimesh parametrat vlerësohen nga moda posterior më mirë se mesatarja posterior.

Laird (1978) vlerësoi qelizat probabilitare duke përdorur vlerësimin empirik Baysian në modelin loglinear. Supozohen priorët me shpërndarje uniforme të përshtatshme për rezultatet e parametrave kryesorë dhe shpërndarje $N(0,\sigma^2)$ të pavarura për parametrat bashkëveprues. Kështu për lehtësi veprimesh parametrat loglinear vlerësohen nga moda e tyre posterior dhe këto posterior mode të përftuar u përdorën në formulën loglineare për të përftuar vlerësues të qelizave probabilitare. Aspekti Bayes-ian empirik kryhet nga zëvendësimi σ^2 me modën e fqinjësisë kufitare, pas integrimit të parametrave loglinear.

Kur $\sigma \to \infty$ vlerësimet konvergjojnë në vëllimin e zgjedhjes
Kur $\sigma \to 0$ vlerësimet konvergjojnë tek vlerësimet e pavarura $\{p_i + p_j\}$.

Vlerat e përftuara kanë të njetën rrjesht dhe shtyllë kufitare si të dhënat e vrojtuara.

Jansen dhe Snijders (1991) në modelin e pavarur përdorin priorët lognormal ose gamma për parametrat në formën multiplicative të modelit. Në mënyrë më të përgjithësuar, Albert (1988) përdori vlerësimin hiearkik për të vlerësuar një model regresi loglinear Puasonian, duke supozuar një prior gamma për mesataren Puasoniane dhe një prior joinformues në parametrat gamma. Tabelat e kontigjencës katrore me kategori të njetë rrjeshtash dhe shtyllash kanë një strukturë të njohur nëpërmjet modeleve që janë të ndryshueshëm për grupe të sigurta të transformimeve të qelizave.

Vaunatsou dhe Smith (1996) analizuan disa stuktura të tilla të tabelës së kontigjencës duke përfshirë :

1. Modelet simetrike
2. Modelet pothuaj simetrike
3. Modelet pothuaj të pavarura për tabelat katrore dhe trekëndshe që rezultojnë kur kategoria korresponduese e qelizës (i,j) është e padallueshme nga ajo e qelizës (j,i).

Ato vlerësuan përshtatjen e modelit duke përdorur distancën dhe duke krahasuar shpërndarjet parashikuese të zgjedhjes që i korrepondojnë vlerave të vëzhguara.

2.9 Përfundime

Në kapitullin 2 matrica e kontigjencës ndërtohet nga tabela e kontigjencës, elementët e së cilës janë të njejtë me të tabelës duke larguar vlerat margjinale. Rangu i matricës përcakton pavarësinë statistikore midis dy karaktereve. Shpërndarja margjinale është e lidhur ngushtë me pavarësinë statistikore midis dy ndryshoreve. Vlerësimet Bayes-iane përdoren për qelizat probabilitare në tabelat e kontigjencës me dy hyrje. Në rastin kur kemi modele me shpërndarje binomiale përdoret shpërndarja prior për të përftuar një proçes prior Dirichle si dhe një vlerësim hiearkik me prior të pavarur beta në parametrat binomial. Në rastin kur kemi modele me shpërndarje multinomiale për vlerësimin e probabiliteteve përdoret prior me të dhëna të pavarura. Ndërsa për vlerësimin e parametrave të modelit loglinear në tabelat e kontigjencës përdoret moda posterior.

Kapitulli 3.

MODELI I REGRESIT LOGJISTIK TË THJESHTE DHE TË SHUMËFISHTË

3.1 Hyrje

Modeli i regresit logjistik për herë të parë u propozua në fund të viteve 60 dhe fillim të viteve 70 si një teknikë alternative e modelit të regresit të përgjithshëm linearë. Në paketat statistikore ai u përfshi në vitin 1980. Përdorimi i tij në fusha të ndryshme kërkimore është gjithnjë e në rritje si në: shkenca sociale (Austin, Yaffee,& Hinkle, 1992, Chuang, 1997), mjeksi (Andrew P. Worth & Mark T.D. Cronin, 2003), arsim (Cabrera, 1994) etj. Në këtë kapitull trajtohet modeli i regresit logjistik të thjeshtë dhe të shumëfishtë, si pjesë përbërse e analizës së të dhënave. Bëhet përshtatja e modelit, kontrolli i rëndësisë së koeficientëve dhe vlerësimi intervalor. Analizohet modeli i regresit logjistik të thjeshtë për parashikimin e pranisë apo mungesës së sëmundjes tek individët. Të dhënat e përdorura janë marrë nga një anketim i bërë në 2016. Rezultatet përftohen nga paketa statistikore SPSS. Në regresin logjistik kur ndyshorja e pavarur është kategorike ajo kodohet (dizenjohet).

3.2 Modeli i Regresit logjistik të thjeshtë

Modelet e Regresit janë pjesë përbërëse e analizës së të dhënave. Ato paraqesin lidhjen midis një variabli të varur dhe disa variablave të pavarur. Gjatë dekadës së fundit modeli i regresit logjistik në shumë fusha studimi është bërë metoda standarte për analizën e të dhënave. Para fillimit të studimit të modelit të regresit logjistik është e rëndësishme të theksojmë se qëllimi i analizës duke përdorur këtë metodë është i njetë me teknikat e tjera të modelimit statistikor. Të gjejmë modelin më të përshtatshëm e të arsyeshëm i cili përshkruan lidhjen midis një variabli të varur dhe një bashkësie me variabla të pavarura. Në modelet e regresit logjistik në krahasim me modelet e regresit linear është se rezultati, variabli i varur është variabël kategorik. Ky ndryshim pasqyrohet dhe në zgjedhjen e një modeli parametrik dhe në supozimet rreth modelit. Pasi ky ndryshim është në rregull, metodat e përdorura në analizën e regresin logjistik ndjekin parimet e përgjithshme të përdorura në regresin linear. Në këtë kapitull trajtohet modeli i regresit logjistik dhe llojet e tij sipas (Hosmer dhe Lemeshow, 2000).

Në punimin Cenaj et al. (2016) kemi ndërtuar dhe analizuar modelin e regresit të thjeshtë logjistik.

Shembull 3.1.

Është bërë një studim në dy zona të ndryshme të Shqipërisë njëra me klimë të ftohtë (Kukës) dhe tjetra me klimë jo të ftohtë. Të dhënat e përdorura paraqiten në tabelën 3.1. Ajo përmban tre ndryshore: moshën, gjininë si dhe praninë ose jo të sëmundjes së reumatizmës për 85 persona nga të cilët 60 banojnë në zonën me klimë të ftohtë dhe 25 në zonën me klimë jo të ftohtë.

Tabela 3.1: Mosha, gjinia prania dhe mungesa e sëmundjes tek 85 persona

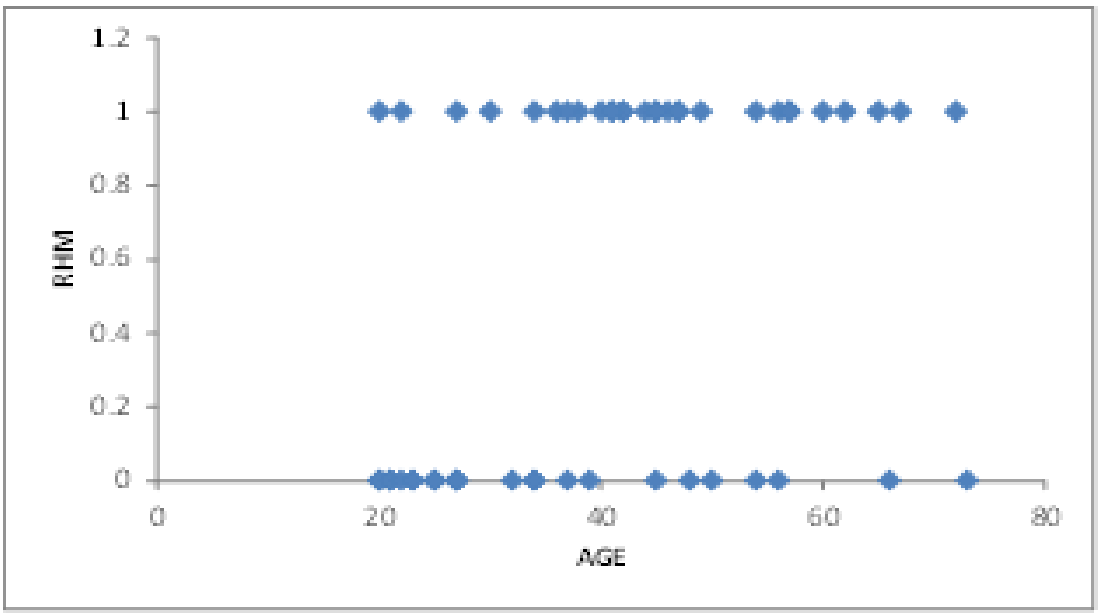
ID	Mosha	Gjinia	Prania ose mosprania e RHM	ID	Mosha	Gjinia	Prania ose mosprania e RHM
1	28	F	1	31	42	M	1
2	30	F	0	32	35	M	0
3	30	M	0	33	40	M	1
4	21	F	0	34	37	F	0
5	18	F	1	35	51	F	1
6	18	F	0	36	49	M	0
7	23	M	0	37	45	F	0
8	37	M	0	38	55	F	1
9	30	M	0	39	52	M	1
10	21	F	1	40	51	M	0
11	24	M	0	41	47	F	0
12	23	F	0	42	45	F	1
13	23	F	0	43	53	F	1
14	25	M	0	44	56	M	1
15	24	M	1	45	45	F	0
16	23	M	0	46	48	F	1
17	22	F	0	47	48	M	0
18	42	F	1	48	55	F	1
19	41	F	1	49	53	M	1
20	41	M	0	50	50	F	1
21	39	F	1	51	69	F	1
22	37	M	0	52	69	F	0
23	36	M	0	53	67	M	1

24	32	F	0	54	61	F	1
25	43	M	1	55	60	F	1
26	42	F	1	56	60	M	1
27	41	F	1	57	58	F	1
28	36	F	0	58	59	F	0
28	33	F	1	59	59	M	1
30	36	M	1	60	60	M	1

Me interes është të shikojmë lidhjen midis moshës dhe pranisë apo mungesës së sëmundjes në zonën me klimë të ftohtë.

Kështu si ndryshore të varur marrim praninë (Y=1) apo mungesën e sëmundjes tek individët (Y=0) dhe si ndryshore të pavarur marrim moshën. Për të njohur më shumë këtë lidhje midis ndryshores të varur sëmundje e reumatizmës (RHM) dhe ndryshores së pavarur (mosha) në figurën 3.1 është paraqitur grafiku me pika.

Figura 3.1: Prania dhe mungesa e sëmundjes për 60 individë



Në figurën 3.1 vihet re se të gjitha pikat bien në një nga dy vijat paralele të cilat përfaqësojnë mungesën dhe praninë e sëmundjes (RHM). Shohim se ka një tendencë që individët të cilët vuajnë nga sëmundja (RHM) të jenë më të mëdhenjë në moshë se individët të cilët nuk vuajnë nga sëmundja.

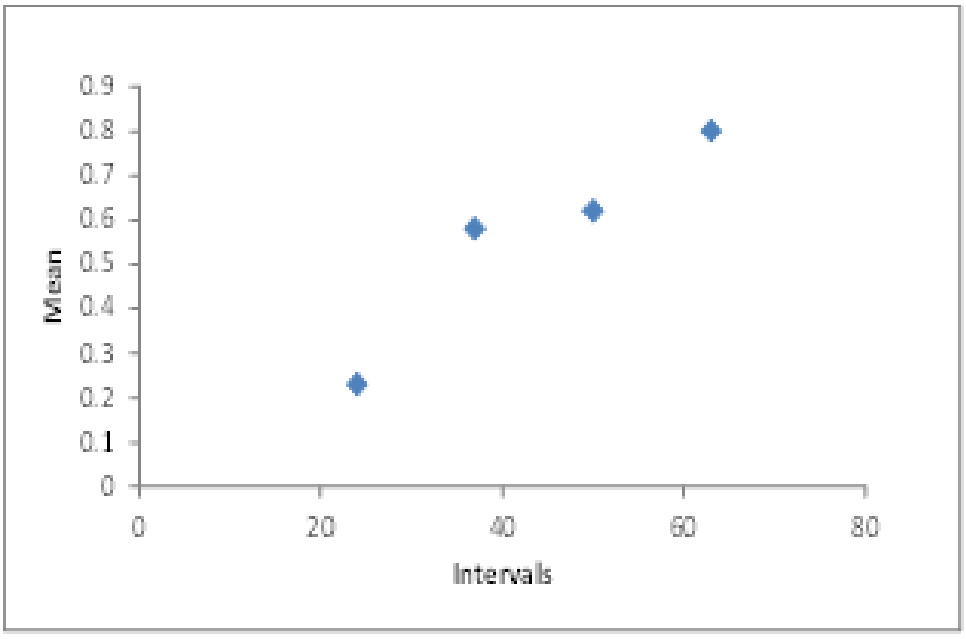
Lidhja funksionale midis sëmundjes së reumatizmës (RHM) dhe moshës është e vështirë të përshkruhet për shkak të ndryshimeve të mëdha të RHM. Për këtë kemi krijuar grupe për variabin e pavarur moshë dhe llogarisim mesataren e RHM për secilin grup.

Tabela 3.2: Denduritë e sëmundjes sipas moshës

	Mungesa e sëmundjes	Prania e sëmundjes	Mesatarja
18-30	13	4	0,23
31-43	7	10	0,58
44-56	6	10	0,62
57-69	2	8	0,8

Në figurën 3.2 ku me anë të pikave paraqet raportin e pranisë së sëmundjes (reumatizmës) në varësi të moshës të kategorizuar në grupe.

Figura 3.2: Raporti i pranisë së sëmundjes në varësi të moshës



Vlerësues sasior në cdo problem regresi është vlera mesatare e variablit të varur, kur njihet vlera e variablit të pavarur. Kjo madhësi quhet pritje matematike e kushtëzuar dhe shënohet

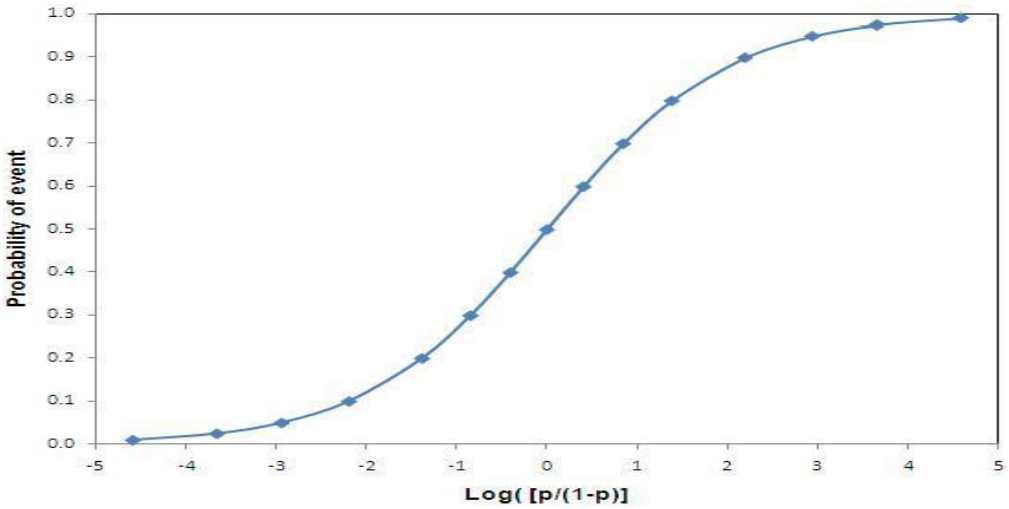
$$E(Y/x),$$

ku Y tregon variablin e varur dhe x tregon një vlerë të variablit të pavarur. Në regresin logjistik mund të përdoret ky shënim vetëm nëse e njohim ligjin probabilitar të Y.

Shpërndarja propabilitare që përdoret për variablat kategorik është shpërndarje logjistike. Për të lehtësuar shënimet bëjmë këtë zëvendësim $\pi(x) = E(Y/x)$ ku Y ka shpërndarje logjistike. Në këto kushte modeli i regresit logjistik ka këtë natyrë

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (3.1)$$

Figura 3.3: Grafiku i funsionit logjistik



Në studimin e regresit logjistik thelbësore është një transformim i $\pi(x)$ i cili është transformim Logit. Ky transformim është përcaktuar në termat e $\pi(x)$, si:

$$g(x) = Logit\big(\pi(x)\big) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

$$= \beta_0 + \beta_1 x$$

Rëndësia e këtij transformimi është se $g(x)$ shqyrtohet si një model i regresit linear. Logit $\pi(x)$, është funksion linear i parametrave i cili mund të jetë i vazhdueshëm dhe me vlera reale të varura nga madhësia e x.

Dallim i rëndësishëm midis modelit të regresit linear dhe modelit të regresit logjistik është shpërndarja e kushtëzuar e variablit të varur.

Më poshtë jepet tabela e cila bën një krahasim midis modelit të regresit linear dhe modelit të regresit logjistik.

Tabela 3.3: Përshkrimi i modelit të regresit logjistik dhe regresit linear

	Regresi Linear	Regresi logjistik
Modeli	$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$	$logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
Përgjigjia (variabli i varur)	Ndryshore të vazhdueshëm	Ndryshore kategorikë
Variabla të pavarur	Të vazhdueshëm ose diskret	Të vazhdueshëm, diskretë, cilësor
Vlerësimi i koeficientëve	Metoda e katrorëve më të vegjël	Metoda e përgjasisë maksimale
Funksioni link	$\eta = g(E(Y_i)) = E(Y_i)$	Logit (π)
Shpërndarja	Variablat përgjigjie janë të pavarur dhe variancë konstante	Variablat përgjigje janë të pavarura dhe secila ka një shpërndarje Bernuli dhe probabiliteti i ngjarjes është i varur nga variablat e pavarur

Në modelin e regresit logjistik kur vlera e një variabli të varur kategorik me dy vlera mund të shprehet $y = E(Y/x) + \varepsilon$, ku madhësia ε

në modelin e regresit logjistik mund të ketë një ose dy vlera të mundshme.

Në qoftë se $y = 1$ atëherë $\varepsilon = 1 - \pi(x)$ me probabilitet $\pi(x)$

Në qoftë se $y = 0$ atëherë $\varepsilon = -\pi(x)$ me probabilitet $1 - \pi(x)$

Kështu, ε ka shpërndarje me mesatare zero dhe variancë të barabartë me $\pi(x)[1 - \pi(x)]$. Në këtë kushte shpërndarja e kushtëzuar e një variabli të varurur ndjek një shpërndarje binomiale me një probabilitet të dhënë nga mesatarja e kushtëzuar $\pi(x)$.

Në analizën e regresit logjistik kur variabli i varur është kategorik plotësohen këto kushte:

1. Mesatarja e kushtëzuar e ekuacionit të regresit merr vlera nga 0 në 1.
2. Shpërndarja e termave të gabimeve është shpërndarje binomiale

3.2.1 Përshtatja e Modelit të Regresit Logjistik

Supozojmë se kemi një zgjedhje me n vrojtime të pavarura të ciftit $(x_i, y_i), i = 1,2, \dots, n$ ku y_i tregon vlerën e variablit të varur (kategorik) dhe x_i është vlera e $i-te$ e variablit të pavarur. Vec kësaj, supozojmë se variabli i varur (kategorik) merr vetëm dy vlera 0 dhe 1, të cilat përfaqësojnë mungesën apo praninë e një karakteristike në subjektet apo individët.

Për të përshtatur modelin e regresit logjistik tek ekuacioni (3.1) për një bashkësi të dhënash kërkohet të përcaktohen vlerat e panjohura të parametrave β_0, β_1. Në modelin e regresit linear, metoda që përdoret për vlerësimin e paramertave të panjohur është metoda e katrorve më të vegjel dhe minimizimi i shumës së katrorve nëpërmjet funksionit të përgjasisë maksimale.

Gjithashtu, në modelin e regresit logjistik për të vlerësuar parametrat përdoret funksioni i përgjasisë maksimale. Fillimisht ndërtohet funksioni i pergjasisë me qëllim zbatimin e metodës së përgjasisë maksimale. Funksion i cili shpreh probabilitetin e të dhënave të vrojtuara si një funksion të parametrave të panjohur. Vlerësuesit e përgjasisë maksimale të këtyre parametrave zgjidhen ato vlera që të maksimizojnë këtë funksion. Në këtë mënyrë, vlerësuesit e parametrave janë më të përafërta me të dhënat e vrojtuara.

Nëse Y merr vetëm dy vlera 0 ose 1 atëherë shprehja për $\pi(x)$ e dhënë në ekuacionin (3.1) siguron (për vektorin e parametrave $\beta = (\beta_0, \beta_1)$, probabilitetin e kushtëzuar që Y të marrë vlerën 1, për x e dhënë e cila shënohet si $P(Y = 1/x)$. Rrjedhimisht madhësia $1 - \pi(x)$ jep probabilitetin e kushtëzuar që Y të marr vlerën 0, kur jepet x.

Kështu që, për çiftin (x_i, y_i),

- kur y_i= 1, për funksionin e përgjasisë kemi $\pi(x_i)$, ndërsa
- kur $y_i = 0$, për funksionin e përgjasisë kemi $1 - \pi(x_i)$, ku madhesia $\pi(x_i)$, tregon vlerën e $\pi(x)$ të llogaritur në x_i .

Dhe vlerat e tij jepen në tabelën e mëposhtme:

Tabela 3.4: Vlerat e Modelit të Regresit Logjistik kur variabli i pavarur është kategorik

Variabli i varur (Y)	Variabli i pavarur (X)	
	$X = 1$	$X = 0$
$Y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$Y = 0$	$1-\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0) = \frac{1}{1+e^{\beta_0}}$
Shuma	1,0	1,0

Meqënëse ndryshoret y_i janë të pavarura sepse vrojtimet janë të pavarura, atëherë funksioni i përgjasisë merret si një prodhim:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)\,]^{1-y_i} \qquad (3.3)$$

Parimi i përgjasisë maksimale përcakton si vlerësues të parametrit β vlerën që maksimizon shprehjen në ekuacionin (3.3). Por, matematikisht më e lehtë është përdorimi i logaritmit të ekuacionit (3.3) si më poshtë:

$$L(\boldsymbol{\beta}) = ln[l(\beta)] = \sum_{i=1}^{n}\{y_i ln[\pi(x_i)] + (1-y_i)ln[1-\pi(x_i)]\}. \quad (3.4)$$

Për të gjetur vlerën e β që maksimizon $L(\boldsymbol{\beta})$ veprohet në këtë mënyrë:

Llogariten derivatet e pjesshme të $L(\boldsymbol{\beta})$ lidhur me β_0, β_1 respektivisht dhe barazohen ato me zero. Këto ekuacione të njohura si ekuacione të përgjasisë janë:

$$\sum\big[[y_i - \pi(x_i)]\big] = 0 \qquad (3.5)$$

dhe

$$\sum x_i\big[[y_i - \pi(x_i)]\big] = 0 \qquad (3.6)$$

Në regresin linear, shprehjet në ekuacionet e përgjasisë janë lineare dhe në këtë mënyrë janë lehtësisht të zgjidhshëm. Ndërsa në regresin logjistik shprehjet në ekuacionet (3.5) dhe (3.6) janë jolineare lidhur me β_0, β_1, kështu në këtë mënyrë kërkojnë metoda speciale për zgjidhjen e tyre.

Këto metoda kanë natyrë përsëritëse dhe janë të programuar në software të posaçme të regresit logjistik. Vlera e $\boldsymbol{\beta}$ që merren si zgjidhje të ekuacioneve (3.5) dhe (3.6) quhet vlerësues i përgjasisë maksimale dhe shënohet $\widehat{\boldsymbol{\beta}}$. Në përgjithësi do të përdoret simboli "^" për vlerësuesit e përgjasisë maksimale të një madhësie të caktar. Shënojmë $\hat{\pi}(x_i)$ vlerësuesin e përgjasisë

makimale të $\pi(x_i)$. Madhësia $\hat{\pi}(x_i)$ jep një vlerësues të probabilitetit të kushtëzuar që *Y* të jetë i barabartë me 1, për *x* e barabartë me x_i . Si e tillë ajo përfaqëson vlerën e parashikuar të modelit të regresit logjistik. Kështu një paraqitje tjetër e ekuacionit (3.5) është:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i).$$

Për të dhënat e shembullit për të bërë vlerësimin e parametrave në modelin e regresit logjistik duke konsideruar si variabël të varur pranimë apo mungesën e sëmundjes së reumatizmës dhe si variabël të pavarur moshën një ndryshore e vazhdueshme.

Tabela 3.5: Përshkrimi i tipareve për bashkësinë e të dhënave

Ndryshoret	Përshkrimi
Prezenca e sëmundjes (reumatizmës)	1-prania e sëmundjes, 0-mungesa e sëmundjes
Mosha	Ndryshore e vazhdueshme

Duke përdorur SPSS marrim këto vlerësime $\widehat{\boldsymbol{\beta}}_0 = -2{,}686$ dhe $\widehat{\boldsymbol{\beta}}_1 = 0{,}068$.

Madhësia $\hat{\pi}(x)$ jepet me këtë barazim:

$$\hat{\pi}(x) = \frac{e^{-2{,}686+0{,}068\cdot mosha}}{1 + e^{-2{,}686+0{,}068\cdot mosha}}$$

Dhe vlerësimi për logit, $\hat{g}(x)$ jepet sipas këtij ekuacioni:

$$logit\big(\pi(x)\big) = \hat{g}(x) = -2{,}686 + 0{,}068 \cdot mosha$$

3.2.2 Kontrolli i rëndësisë së koeficientëve

Në modelin e regresit logjistik pasi vlerësohen koeficientët e rëndësishme është kontrolli i rëndësisë së tyre. Kjo zakonisht përfshin formulimin dhe testimin e një hipoteze statistikore për të përcaktuar nëse variablat e pavarura në modelin janë të lidhura në mënyrë të konsiderueshme (të rëndësishëm statitikisht) me variablat e varur.

Në modelin e regresit logjistik për kontrollin e rëndësisë së parametrave përdoret testi i krahasimit të vlerave të vrojtuara të variablit me vlerat e përafruara të tij sipas modelit të ndërtuar. Krahasim i cili bëhet duke marrë logaritmin e funksionit të përgjasisë (3.4). Për ta kuptuar më mirë këtë krahasim është e nevojshme të mendojmë një vlerë të vrojtuar të variablit të varur dhe një vlerë të parashikuar të tij (vlerësuar) të marrë si rezultat i një modeli të "ngopur". Një model është i "ngopur" nëse ai përmban aq parametra sa variabla të vrojtuar. Një shembull i thjeshtë i një modeli të "ngopur" është një model i regresit linear, kur ka vetëm dy pika të të dhënave.

Krahasimi i vlerave të vrojtuara dhe atyre të parashikuara duke përdorur funksionin e përgjasisë paraqitet si më poshtë:

$$D = -2ln\left[\frac{(pergjasia\ per\ modelin\ e\ pershtatur)}{(pergjasia\ per\ modelin\ e\ ngopur)}\right] \qquad (3.7)$$

Ku, raporti njihet si raport i përgjasive ndërsa testi si testi i raportit të përgjasive.

Duke përdorur ekuacioni (3.4) dhe (3.7) kemi:

$$D = -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\widehat{\pi_i}}{y_i}\right) + (1-y_i)ln\left(\frac{1-\widehat{\pi_i}}{1-y_i}\right)\right] \qquad (3.8)$$

ku $\widehat{\pi_i} = \widehat{\pi}(x_i)$

Statistika D, e përftuar si në (3.8) quhet *deviancë* dhe luan një rol kryesor për të vlerësuar cilësinë e përshtatjes së modelit e quajtur *përputhshmërinë e modelit.*

Në modelin e regresit logjistik devianca luan të njetin rol si shuma e katrorve të gabimit në regresin e përgjithshëm linear e cila tregon se në qoftë se ekuacioni i përftuar i regresit do të shpjegonte plotësisht vrojtimet d.m.th në qoftë se të gjithë pikat e vrojtuara do të ndodheshin në drejtëzën e regresit, atëherë kjo shumë do të ishte 0.

Për më tepër, kur vlerat e variablit të varur janë si në tabelën (3.1) d.m.th ato marrin vetëm dy vlera 0 dhe 1, atëherë modeli i përgjasisë së ngopur është 1.

Specifikisht, nga përkufizimi i një modeli të "ngopur" rrjedh se $\widehat{\pi_i} = y_i$ dhe përgjasia është:

$$l(model\ i\ ngopur) = \prod_{i=1}^{n} y_i{}^{y_i} \times (1-y_i)^{(1-y_i)} = 1.$$

Kështu nga ekuacioni (3.7) deviance (shmagia) është:

$$D = -2\ln(pergjasia\ e\ modelit\ te\ pershtatur). \quad (3.9)$$

Disa paketa software, të tilla si SPSS, SAS gjejnë vlerat e deviancës në (3.9) në vend të log përgjasisë për modelin e përshtatur.

Për të vlerësuar rëndësinë e një variabli të pavarur në modelin e regresit logjistik, krahasohen vlerat e D-së me dhe pa variablat e pavarur në ekuacion. Për shkak të përfshirjes së variablave të pavarur në model sjell ndryshimin e D-së e cila shënohet:

$$G = D(modeli\ pa\ variablin) - D(modeli\ me\ variablin).$$

Ku $D(modeli\ pa\ variablin)$ është analoge me shumën e katrorëve të totalit tek regresi I përgjithshëm linear, ndërsa $D(modeli\ me\ variablin)$ është analoge me shumën e katroreve të gabimit.

Duke përdorur parimin e përgjasisë maksimale statistika G merr trajtën:

$$G = -2ln\left[\frac{(pergjasia\ pa\ variabla)}{(pergjasia\ me\ variabla)}\right] \quad (3.10)$$

Për rastin specifik kur kemi një variabël të vetëm të pavarur, tregohet se vlerësuesi i përgjasisë maksimale për parametrin β_0 është $ln\left({}^{n_1}/_{n_0}\right)$ ku $n_1 = \sum y_i$ dhe $n_0 = \sum(1 - y_i)$

dhe vlera e parashikuar është kostante, ${}^{n_1}/_{n_0}$. Në këtë rast vlera e G është:

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi_i}^{y_i}(1-\widehat{\pi_i})^{(1-y_i)}}\right] \quad (3.11)$$

ose

$$G = 2\{\sum_{i=1}^{n}[y_i\ln(\widehat{\pi_i}) + (1-y_i)\ln(1-\widehat{\pi_i})] - [n_1\ln(n_1) + n_0\ln(n_0) - nln(n)]\} \quad (3.12)$$

Kur $H_0 = \beta_0$ atëherë statistika G ka shpërndarje Hi-katror (χ^2) me një shkallë lirie.

Një statistikë tjetër është statistika Wald e cila përftohet duke krahasuar vlerësuesin e përgjasisë maksimale $\widehat{\beta_1}$ të β_1 me një vlerësues të gabimit standart të tij.

Pra ajo përdoret për të kontrolluar nëse një ndryshore e pavarur X është e rëndësishme për shpjegimin e ndryshores së varur Y.

Statistika Wald jepet me këtë formulë:

$$W = \frac{\widehat{\beta_1}}{\widehat{SE}\,(\widehat{\beta_1})}$$

dhe ka shpërndarje normale standarte. Për një zgjedhje me vëllim të vogël vlera e kësaj statistike llogaritet lehtë, ndërsa për zgjedhje me vëllim të madh, zbatimi i këtij testi mund të japë rezultate të gabuara, domethënë ne mund të largojmë një ndryshore të pavarur (shpjeguese) nga modeli ndërsa ajo mund të jetë e rëndësishme. Kjo gje u vu re nga Hauck dhe Donner (1977) shqyrtuan performancën e testit Wald dhe treguan se në disa raste ajo nuk vlen, për të hedhur poshtë hipotezën zero kur koeficienti ishte i rëndësishëm.

Goodness of fit of the model përdoret për të kontrolluar vlefshmërinë e modelit, pra, sa 'mirë' modeli i përftuar përshtatet tek të dhënat, ndryshe tregon sa afër janë vlerat e parashikuara të ndryshores së varur Y me vlerat reale të saj.

Gjithashtu llogariten dhe dy tregues të tjerë si: Cox Snell R^2, Nagelkerke R^2 të përcaktuara nga Cox dhe Snell (Cox, D.R. & Snell, E.J.,1989) dhe Nagelkerke (Nagelkerke, N. J. D., 1991). Interpretimi është i ngjashëm me koeficientin e përcaktueshmërisë tek regresi linear.

Vlerat e këtyre statistikave nuk janë të njëjta, sepse janë llogaritur në shkallë të ndryshme.

3.2.3 Vlerësimi i intervalit të besimit

Duke u mbështetur në statistikën Wald ndërtohen edhe intervalet e besimit për nivel rëndësie α të dhënë, të parametrave të modelit si m poshtë:

$$\widehat{\beta_1} \pm \boldsymbol{z_{1-\alpha/2}}\widehat{SE}\,(\widehat{\beta_1}) \qquad (3.13)$$

dhe

$$\widehat{\beta_0} \pm \boldsymbol{z_{1-\alpha/2}}\widehat{SE}\,(\widehat{\beta_0}) \qquad (3.14)$$

Logit është forma lineare e modelit të regresit logjistik kështu që, ka më shumë ngjashmëri me modelin e regresit linear. Vlerësuesi për modelin logit është:

$$\widehat{g}(x) = \widehat{\beta_0} + \widehat{\beta_1}x \qquad (3.15)$$

Vlerësimi i variancës për këtë vlerësues jepet nga barazimi:

$$\widehat{Var}[\hat{g}(x)] = \widehat{Var}(\widehat{\beta_0}) + x^2\widehat{Var}(\widehat{\beta_1}) + 2x\widehat{Cov}(\widehat{\beta_0}, \widehat{\beta_1}) \qquad (3.16)$$

Intervali i besimit i bazuar në statistiken Wald për modelin logit jepet si më poshtë:

$$\widehat{g}(x) \pm \boldsymbol{z_{1-\alpha/2}}\widehat{SE}\,[\widehat{g}(x)] \qquad (3.17)$$

Ku $\widehat{SE}\,[\widehat{g}(x)] = \sqrt{\widehat{Var}[\hat{g}(x)]}$

$$= \sqrt{\widehat{Var}(\widehat{\beta_0}) + x^2\widehat{Var}(\widehat{\beta_1}) + 2x\widehat{Cov}(\widehat{\beta_0}, \widehat{\beta_1})} \qquad (3.18)$$

3.2 Modeli i Regresit logjistik të shumëfishtë

Në këtë pjesë do të trajtojmë modelin e regresit logjistik të shumëfishtë i cili ndërtohet në rastin kur kemi një variabël të varur që merr vetëm dy vlera dhe $p-$ variabla të pavarur.

Supozohet se kemi variablin e varur $Y = \{0,1\}$ dhe p −variabla të pavarur $x_1, x_2, \dots, x_p$, $p \geq 2$, shënojmë $x' = (x_1, x_2, \dots, x_p)$ vektori i cili ka variablat e pavarur dhe $P(Y = 1/x) = \pi(x)$.

Në rastin kur variablat e pavarura marrin vlera në intervale atëherë modeli Logit për modelin e regresit logjistik të shumëfishtë jepet me këtë ekuacion:

$$g(x) = Logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (3.19)$$

Ku modeli i regresit logjistik është:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \qquad (3.\ 20)$$

Në qoftë se disa nga variablat e pavarur janë diskrete, variabla me shkallë nominale (të cilat marin vlera jo sipas nje renditje te caktuar) të tilla si: raca, seksi etj, ajo është e papërshtatshme që të përfshijë ato në modelin sikur të ishin ndryshore me shkallë interval. Numrat e përdorura për të përfaqësuar nivelet e ndryshme të këtyre variablave me shkallë nominale janë thjesht identifikues domethënë janë variabla kategorik ata do të konsiderohen si variabla të dizenjuar (koduar).

Në përgjithësi nëse variabli me shkallë nominal ka k-vlera të mundshme, atëherë ndërtohen (k-1) variabla të dizenjuar (koduar). Për ta demostruar këtë, pra ndërtimin e variablit të dizenjuar supozoj se variabli i pavarur x_j ka k_j vlera, kështu për variablin x_j do të ndërtohen $k_j - 1$

variabla të dizenjuar të cilat do ti shënojmë me D_{jl} me koeficient përkatës β_{jl} ku $l = 1,2, \dots, k_j - 1$. Kështu modeli logit i paraqitur më sipër (3.19) me p variabla, ku variabli i $j-te$ është kategorik merr këtë trajtë:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p. \qquad (3.21)$$

3.3.1 Përshtatja e Modelit të Regresit Logjistik të Shumëfishtë

Supozohet se kemi një zgjedhje me n vrojtime të pavarura (x_i, y_i), $y_i = 1,2, \dots, n$. Përshtatja e modelit të regresit logjistik të shumëfishtë kërkon gjetjen e vlerësuesve të parametrave në vektorin $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$. Ndërsa si metodë e vlerësimit do të jetë metoda e përgjasisë maksimale. Funksioni i përgjasisë është pothuajse identik me ekuacionin e dhënë (3.3) por me të vetmin ndryshim se $\pi(x)$ është e përcaktuar sipas ekuacionit (3.20). Janë $p + 1$ ekuacione përgjasie, të cilët përftohen duke llogaritur derivatet e pjesshme të funksionit të përgjasise në lidhje me p+1 koeficientet respective. Ekuacionet e përgjasisë të përftuar në këtë mënyrë janë:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

dhe

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0$$

për $j = 1,2, \dots, p$.

Zgjidhja e ekuacioneve të përgjasisë bëhet nëpërmjet programeve të posaçëm që gjenden thuajse në të gjitha paketat statistikore. Zgjidhjen e këtyre ekuacioneve do ta shënojmë $\widehat{\boldsymbol{\beta}}$. Për modelin e regresit logjistik të shumëfishte vlerat e përshtatura janë $\hat{\pi}(x_i)$, vlera në ekuacionin (3.20) llogariten duke përdorur $\widehat{\boldsymbol{\beta}}$ dhe x_i.

Metoda për vlerësimin e variancës dhe kovariancës së vlerësueseve të koeficentëve ndjek parimin e përgjasisë maksimale. Vlerësuesit merren nga matricat e derivateve të pjesshme e rendit të dytë të funksionit log të përgjasisë. Derivatet e pjesshme të rendit të dytë jepen nga barazimet e mëposhtme:

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j{}^2} = -\sum_{i=1}^{n} x_{ij}^2\, \pi_i(1-\pi_i) \quad (3.22)$$

dhe

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_l} = -\sum_{i=1}^{n} x_{ij}\, x_{il} \pi_i(1-\pi_i) \quad (3.23)$$

Për, $= 1,2, \dots, p$, ku $\pi_i = \pi(x_i)$.

Matrica e cila ka si element, elementët negativ të cilët meren nga ekuacionet (3.22) dhe (3.23) do të quhet matrica e informacionit të vrojtuar. E cila shënohet $I(\beta)$ dhe ka përmasa (p + 1) × (p + 1). Variancat dhe kovariancat e vlerësuesve të koeficientëve gjenden nga matrica e anasjellë e matricës së informacionit të vrojtuar pra $Var(\beta) = I^{-1}(\beta)$.

Elementin e j-të të diagonals së kësaj matrice do ta shënojmë me $Var(\beta_j)$ e cila është variance e $\widehat{\beta_j}$. Shënojmë $Cov(\beta_j, \beta_l)$ elementët jashtë diagonales kovariancën midis $\widehat{\beta_j}$ dhe $\widehat{\beta_l}$. Vlerësuesit e variancës dhe kovariancës shënohen me $\widehat{Var}(\hat{\beta})$ të cilët merren duke vlerësuar $Var(\beta)$ për $\hat{\beta}$. Shënojmë me $\widehat{Var}(\widehat{\beta_j})$ dhe $\widehat{Cov}(\widehat{\beta_j}, \widehat{\beta_l})$, $j, l = 1,2, \dots, p$ elementët e matricës.

Vlerësuesi gabimit standart për vlerësuesit e koeficientëve jepet si:

$$\widehat{SE}(\widehat{\beta_j}) = \left[\widehat{Var}(\widehat{\beta_j})\right]^{1/2} \quad (3.24)$$

për $j = 1,2, \dots, p$.

Matrica e informacionit të vrojtuar mund të paraqitet në trajtë tjetër matricore si

$\widehat{\boldsymbol{I}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X'VX}$, ku X është matricë e cila përmban të dhëna për secilin vrojtim dhe ka përmasa $n \times (p+1)$ ndërsa V është matrica me përmasa $n \times n$ ku si element të saj ka element të kësaj natyre $\widehat{\pi_i}(1-\widehat{\pi_i})$.

Kështu matricat X dhe V kanë këtë formë:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

$$V = \begin{pmatrix} \widehat{\pi_1}(1-\widehat{\pi_1}) & 0 & \dots & 0 \\ 0 & \widehat{\pi_2}(1-\widehat{\pi_2}) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & & \widehat{\pi_n}(1-\widehat{\pi_n}) \end{pmatrix}$$

3.3.2 Kontrolli i Rëndësisë së Modelit

Ashtu si rastin e modelit të regresit logjistik të thjeshtë, dhe në modelin e regresit logjistik të shumëfishtë vlerësimi për modelin është vlerësimi i rëndësisë së variablave në model.

Kriteri i raportit të përgjasisë për rëndësinë e përgjithshme të p koeficientëve të variablave në model (të pavarur) zbatohet si në rastin e modelit të regresit logjistik të thjeshtë.

Kriter i cili bazohet në statistikën G të paraqitur në ekuacionin (3.10) por me një ndryshim që vlerat e përshtatura $\hat{\pi}$ bazohen në vektorin $\hat{\beta}$ i cili ka $(p+1)$ parametra.

Lidhur me hipotezën zero kur të gjithë koeficientët janë të barabartë me zero pra:

$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$ shpërndarja e statistikës G eshtë Hi katror me p shkallë lirie. Përdoret statistika Wald për të kontrolluar rëndësinë e secilit prej koeficientëve të pavarur

$$W_j = \hat{\beta}_j / \widehat{SE}(\hat{\beta}_j)$$

Secila statistikë W_j ka ligj normal të normuar. Ndërkohë që reduktohet numri i parametrave qëllimi kryesor është është të merret modeli më i përshtatshëm.

Në vazhdim është që të përshtatet modeli i cili përmban vetëm variablat e rëndësishëm dhe të krahasohet ai me modelin e fillimit i cili i përmban të gjitha variablat. Për kontrollin e hipotezës kriteri Wald analog i modelit të shumëfishtë përftohet në trajtë matricore si:

$$W = \widehat{\beta}' \left[\widehat{Var}(\hat{\beta}_j)\right]^{-1} \hat{\beta}$$

$$= \widehat{\beta}' (\boldsymbol{X'VX}) \, \hat{\beta}.$$

W ka shpërndarje Hi katror me $(p+1)$ shkallë lirie.

3.4 Përfundime

Në këtë kapitull analizohet modeli i regresit logjistik të thjeshtë dhe të shumëfishtë, përshtatja e modelit, kontrolli i rëndësisë së koeficientëve dhe vlerësimi i intervalit të besimit për këto modele. Për parashikimin e probabilitetit të pranisë apo mungesës së sëmundjes tek individët u ndërtua modeli i regresit logjistik të thjeshtë $logit(\pi(x)) = -2{,}686 + 0{,}068 \cdot mosha$, të dhënat e përdorura janë marrë nga një anketim i bërë në 2016 në zonën e Kukësit. Në modelin e regresit logjistik kur ndryshorja e pavarur është kategorike, ajo kodohet.

KAPITULLI 4

MODELIMI I TË DHËNAVE ME REGRESIN LOGJISTIK TË SHUMËFISHTË

4.1 Hyrje

Në këtë kapitull realizohet modelimi i të dhënave me regresin logjistik të thjeshtë dhe të shumëfishtë. Modelimi parashikon probabilitetin e mos përfshirjes në Skemën e Sigurimeve Shoqërore të një individi të punësuar. Për ndërtimin e këtij modeli të dhënat janë marrë nga baza e të dhënave e The World Bank në një raport të paraqitur më 2014 mbi anketimin "Matja e standarteve për nivelin e jetesës në Shqipëri" të realizuar më 2012. Analiza e regresit logjistik është kryer në disa hapa. Në fillim është marrë vetëm me një ndryshore dhe më pas janë shtuar ndryshoret e tjera duke përmirësuar modelin. i gjithë zbatimi është realizuar duke përdorur paketën statistikore SPPS.

4.2 Historik i shkurtër mbi zhvillimin e sigurimeve shoqërore në Shqipëri

Përballjet me shumë realitete kanë parashtruar nevojën për mbrojtje shoqërore, duke e përkufizuar atë si një aspiratë jo vetëm të një klase shoqërore, të një grupi të caktuar individësh, të një kategorie të caktuar profesionale, por si një e drejtë njerëzore. Skema e sigurimeve në Shqipëri është e mbështetur në sistemin "pay as you go", sipas të cilit pagesat aktuale për pensionistët financohen nga kontributet aktuale brenda sistemit. Kjo krijon një lidhje të drejtpërdrejtë midis numrit të personave që kontribuojnë në skemë, sasisë së parave që ata paguajnë dhe pensioneve të përfituara. Sa më i lartë ky raport, aq më mirë është. Mënyra e menaxhimit ndër vite e skemës së sigurimeve shoqërore dhe karakteristikat e saj kanë qenë të ndryshme para dhe pas viteve '90. Në Periudhën e dytë, e cila paraqet shumë interes nga pikëpamja e analizës së pjesëmarrësve në skemën e sigurimeve i përket viteve 1991-2009.(sipas Insitutit të Sigurimeve Shoqërore). Zhvillimet e vrullshme në fillim të viteve '90 u shoqëruan me mbylljen e shumë fabrikave, rritje të papunësisë, përpjekje për krijimin e një sistemi të ri ekonomik, i cili në fillesat e tij u karakterizua nga informalitet i lartë, shkaqe këto të gjitha që sollën kolaps në numrin e kontribuesve. Për të neutralizuar këtë rënie, në vitin 1993, me hartimin e ligjit për Sigurimet Shoqërore, u vendos një normë e lartë për kontributet, në masën 42.5%.

Pas vitit 2000 tregu vazhdonte të karakterizohej nga mungesë informacioni mbi efektivitetin e veprimtarisë ekonomike, kontroll jo i plotë i subjekteve dhe numrit të të punësuarve në to, pagat reale, informalitet i lartë sidomos në industrinë e ndërtimit, mospërputhje të numrit të firmave të licencuara dhe atyre që derdhnin kontribute, përfshirje gjithmonë e më e ulët e punëtorëve në sektorin e bujqësisë, etj. Në këtë kuadër, me qëllim rritjen e numrit të personave të përfshirë në skemë, në vitin 2002, norma e kontributeve u ndryshua nga 42.5% në 38.5%.

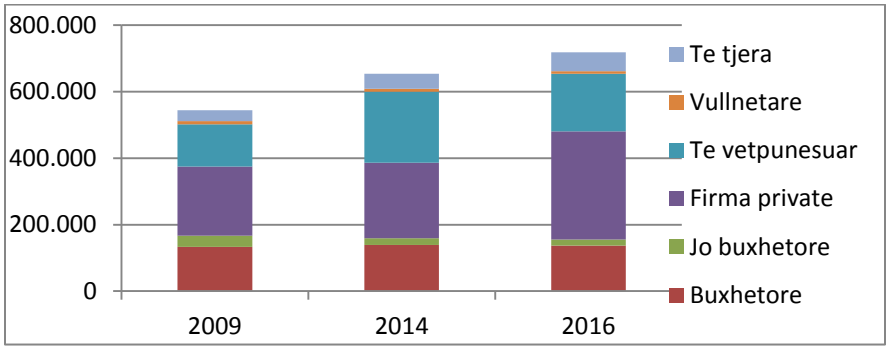
Në vitin 2003 efekti i kësaj reforme nuk u ndie pasi numri i kontribuesve mbeti pothuajse në të njëjtat nivele, ndërsa në vitin 2004, filloi të jepte shenjat e para duke u rritur me 1.7%, në vitin 2005 me 3.4% e në vitin 2006 me 11.2%. Rritjen më të madhe po e shënonin të vetpunësuarit dhe firmat private. Numri total i kontribuesve ka shënuar numrin më të madh në vitin 2016 me mesatarisht 718286 kontribues, Nga ana tjetër, numri më i ulët i kontribuesve është shënuar në vitin 2009, me rreth 544591 kontribues, (sipas Insitutit të Sigurimeve Shoqërore).

Tabela 4.1: Numri i kontribuesve në skemën e Sigurimeve Shoqërore sipas sektorit të punësimit

	Buxhetore	**Jo buxhetore**	**Firma private**	**Të vetpunesuar**	**Vullnetare**	**Të tjera**
2009	133,155	34,219	207,798	127,151	8,999	33,269
2014	138,651	20,552	226,886	213,896	8,534	46,044
2016	137,012	18,389	325,303	173,628	7,323	56,631

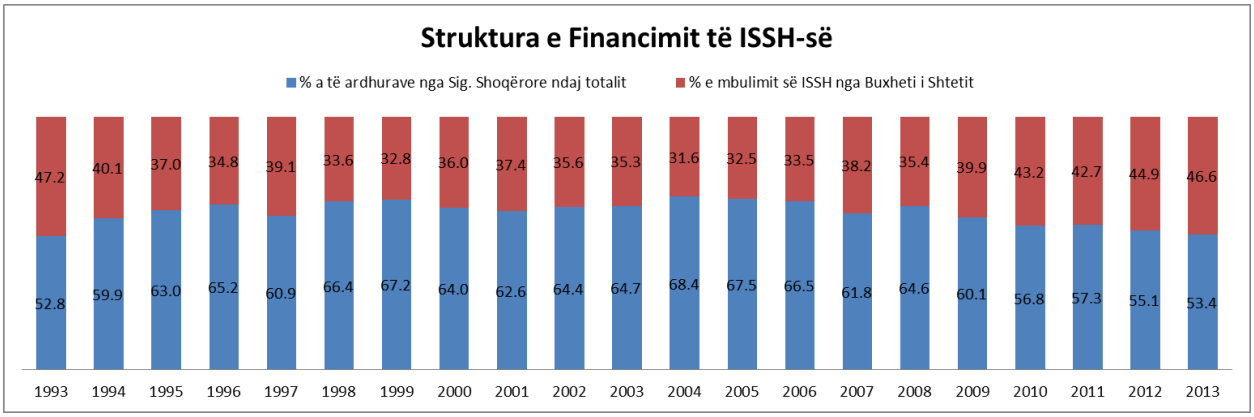
Grafikisht të dhënat e mësipërme i paraqesim si më poshtë ku vemë re se numri individëve të përfshirë në Skemën e Sigurimeve Shoqërore në sektorin privat ka ardhur në rritje të vazhdueshme. Ndërsa numri i individëve të përfshirë në skemë në sektorin Jo buxhetor ka ardhur në një ulje të vazhdueshme. Grafikisht këtë mund ta paraqesim si më poshtë:

Figura 4.1: Kontribuesit në skemën e sigurimeve shoqërore sipas sektorve



Skema e sigurimeve thith një pjesë jo të vogël të fondeve buxhetore. Në paraqitjen grafike të realizuar në excel, jepen të ardhurat nga kontributet dhe shpenzimet për sigurime shoqërore për periudhën 1993-2013.

Figura 4.2: Struktura e financimit të ISSH.

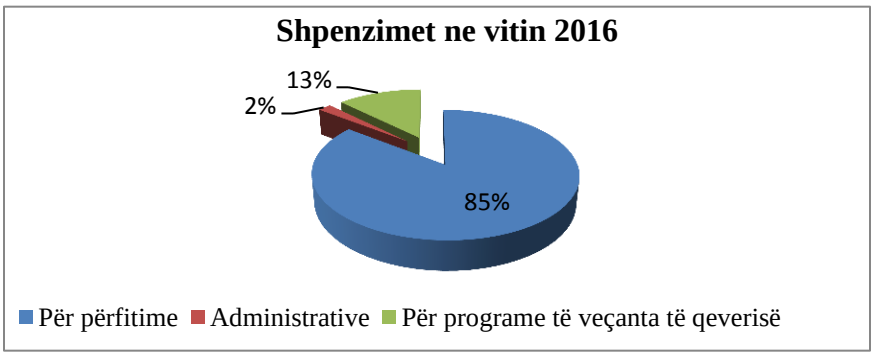


Vihet re nga grafiku i mësipërm një rritje e vazhdueshme e të ardhurave nga kontributet, sidomos në vitet 1993-1996. Por me kalimin e viteve, hendeku midis shpenzimeve dhe të ardhurave ka ardhur duke u thelluar dhe varësia nga Buxheti i Shtetit është kthyer në domosdoshmëri. E vlerësuar në kontekstin se Skema e Sigurimeve Shoqërore është një skemë e mbrojtjes sociale, por funksionon si skemë kontributive, pagesa e kontributit të sigurimeve shoqërore gjatë gjithë periudhës së punësimit të një individi, është "primi" i kontratës që individi ka me Institutin e Sigurimeve Shoqërore. Kështu struktura e shpenzimeve jepet sipas tabelës 4.2

Tabela 4.2: Struktura e shpenzimeve në vitin 2016

Shpenzime	**Përqindja**
Për përfitime	85,5
Administrative	2,0
Për programe të veçanta të qeverisë	12,5

Figura 4.3: Shpenzimet në vitin 2016



Numri mesatar i kontribuesve në vit ka patur luhatjet e tij nga viti 2011 deri në vitin 2016, ku

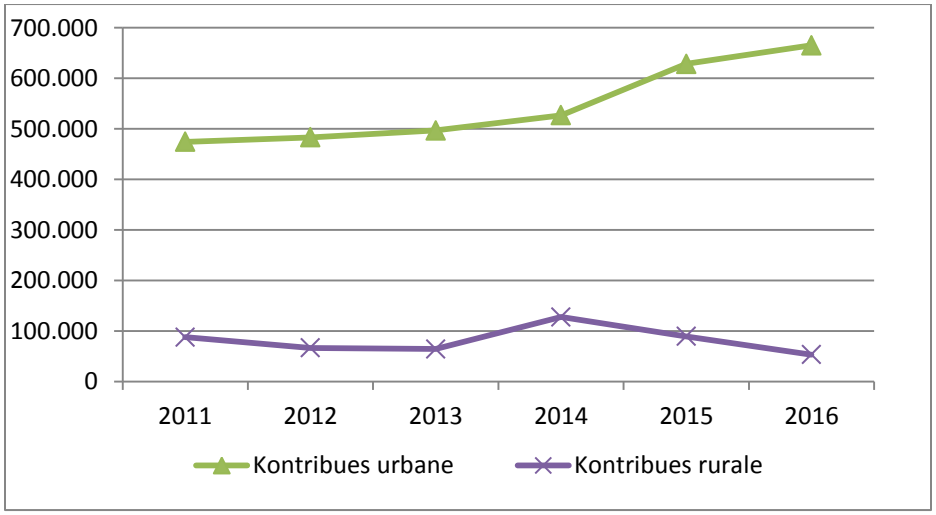
ndikim të konsiderueshëm kanë patur ulje-ngritjet e numrit të të vetëpunësuarve në bujqësi. Në vitin 2016 u shënua numri më i lartë i kontribuesve prej 718286 kontribues, ndërsa në vitin 2012 u shënua numri më i ulët i kontribuesve me vetëm 549720 kontribues. Njëkohësisht, viti 2016 është viti me rritjen vjetore më të lartë të kontribuesve me rreth 9%, ndërsa viti 2012 është viti me rënien vjetore më të theksuar të kontribuesve me minus 11.4%.

Tabela 4.3: Kontribues të sigurimeve shoqërore në vite

	2011	2012	2013	2014	2015	2016
Konribues në total	562,146	549,720	561,169	654,563	718,070	718286
Kontribues urbane	474,351	483,100	496,895	526,835	628,543	665118
Kontribues rurale	87,795	66,620	64,274	127,728	89,527	53168

Në Skemën e Sigurimeve Shoqërore numrin mesatar më të madh të kontribuesve në vit e zënë kontribuesit në zona urbane. Përgjatë viteve 2011-2016 numri mesatar i kontribuesve urbanë ka zënë rreth 63% të totalit të kontribuesve, ndërsa pjesa tjetër prej 37% i përket peshës së kontribuesve ruralë. Numri i kontribuesve urbanë shënoi vlerën më të ulët të tij në vitin 2011 me vetëm 474351 kontribues. Nga ana tjetër, numri më i lartë i kontribuesve ruralë u shënua në vitin 2014, ku kishim 127728 kontribues dhe numri më i ulët i kontribuesve ruralë u shënua në vitin 2013, ku kishim vetëm 53168 kontribues. Grafikisht këto të dhëna paraqiten ne figurën 4.4:

Figura 4.4: Numri i kontribuesve urbanë e rurale në vite



4.3 Të dhënat

Do të përdorim një zgjedhje, të marrë nga baza e të dhënave i The World Bank në një anketimi të bërë "Matja e standarteve të jetesë në Shqipëri 2012" raport i paraqitur më 13 mars 2014. Këto të dhëna do ti përdorim për të parashikuar probabilitetin e mos përfshirjes së një individi të punësuar në Skemën e Sigurimeve Shoqërore. Kjo bashkësi të dhënash përbëhet nga $n = 3310$ të cilët janë zgjedhur në dy raunde. Në raundin e parë, 834 Njësitë Përzgjedhëse Primare (PSU) janë zgjedhur rastësisht për të përfaqësuar të gjithë territorin e vendit. Pastaj, 8 familje për secilën PSU janë zgjedhur për t'u intervistuar në raundin e dytë përmes një procedure. Për të trajtuar rastet e mos përgjigjes ose të asnjë kontakt tjetër, 4 familje për secilën PSU u zgjodhën si zëvendësues që sigurojnë objektivin e 3310 pyetësorëve të plotësuar pranë familjeve. Fusha gjeografike të analizës përfshijnë 12 prefekturat e Shqipërisë, zonat urbane dhe rurale si dhe të dy gjinitë. Nga të dhënat e anketimit po të bëjmë një klasifikim përsa i përket përfshirjes ose jo në skemë të individëve marrim tabelën 4.4:

Tabela 4.4: Shpërndarja e individëve sipas pjesmarrjes në skemën e sigurimeve shoqërore

		Denduria	**% për dendurinë**
Pjesmarrja në sigurime shoqërore	Po	2331	70.4
	Jo	979	29.6
	Total	3310	100.0

Nga baza e të dhënave është përdorur një ndryshore e varur dhe 5 ndryshore të pavarura. Ndryshorja e varur është ndryshore e cila merr dy vlera dhe tregon nëse një individ është pjesë e skemës së sigurimeve shoqërore ose jo. Variablat e pavarura janë: vitet e shkollimit, lloji i punësimit, mosha, gjinia dhe zona. Ndryshoret e kësaj bashkësie të dhënash jepen nga tabela 4.5.

Tabela 4.5: Përshkrimi i tipareve për bashkësinë e të dhënave

Ndryshoret	**Pëshkrimi**
Përfshirja në skemën e sigurimeve shoqërore	0→ përfshihet, 1→nuk përfshihet
Mosha	Ndryshore e vazhdueshme
Gjinia	0→ mashkull, 1→ femër
Vitet e shkollimit	Ndryshore e vazhdueshme
Lloji i punësimit	1→ Sektori publik, 2→ Kompani private 3→ Program i punës publike 4→ Ndërmarrje shtetërore 5→ Në OJQ ose organizata humanitare 6→ Vetëpunësuar
Zona	1→ Qëndrore 2→ Bregdetare 3→ Malore 4→ Tirana

4.4 Modelimi i të dhënave me regresin logjistik të shumëfishtë

Në fillim llogarisim tabelat e kontigjencës për ndryshoren e pavarur Gjinia dhe ndryshoren e varur përfshirja në Skemën e Sigurimeve Shoqërore për të treguar nëse ka ndryshim mes femrave dhe meshkujve dhe përfshirjes në Skemën e Sigurimeve Shoqërore.
Nga tabela e kontigjencës vemë re se 35,38% janë femra nga të intervistuarit dhe 64,62% janë meshkuj dhe nga këto 84.5% e femrave janë të përfshira në skemë dhe 62.7% e meshkujve janë të përfshirë në skemë. Pra pritet që një femër të ketë më shumë shanse se një mashkull për tu

përfshirë në skemë, duke marrë parasysh që regresi logjistik na jep probabilitetin e ngjarjes së koduar me 1(Y=1 në intepretimin tona do ti kushtojme rëndesi mos përfshirjes në skemë)

Tabela 4.6: Tabela e kontigjencës për shpërndarjen e individëvë të përfshirë në skemë sipas gjinisë

			Gjinia		Totali
			Mashkull	Femër	
Përfshirja në skemë	**Po**	Denduri	1341	990	2331
		% Për Gjinia	62.7%	84.5%	70.4%
	Jo	Denduri	798	181	979
		% për Gjinia	37.3%	15.5%	29.6%
Totail		Denduri	2139	1171	3310
		% Për Gjinia	100.0%	100.0%	100.0%

Nga tabela 4.6 vihet re se numri femra të përfshira në skemën e sigurimeve shoqërore është në numër më i madh se ai meshkujve.

Njehsojmë raportin e mundësive (Odds) si më poshtë:

$$Odds = \frac{P(A)}{1 - P(A)} = \frac{prob\ 1 - ve}{prob\ 0 - ve}$$

Për ndryshoren me dy vlera në modelin e regresit logjistik modelojmë logaritmin natyral të raportit të gjasave, $logit(\pi)$

$$logit(\pi) = \ln(Odds) = ln\left(\frac{\pi}{1 - \pi}\right)$$

$$= ln\left(\frac{prob\ 1 - ve}{prob\ 0 - ve}\right)$$

Ai bazuar në të dhënat e tabelës 4.6 raporti i gjasave është 0.419. Ky raport tregon se nëse zgjedhim rastësisht një individ ai është 0.419 herë më i favorshëm për të mos qenë i përfshirë sesa i përfshirë në skemën e Sigurimeve Shoqërore.

4.4.1 Rasti i parë (Varësia midis dy variablave kategorik)

Raporti i gjasave përkatësisisht për femra dhe meshkuj jepet me anë të barazimeve:

$$Odds_F = \frac{joperfF}{perF} = \frac{181}{990} = 0.1828$$

dhe

$$Odds_M = \frac{joperfM}{perM} = \frac{798}{1341} = 0.5951$$

Raporti i dy gjasave tregon $\frac{Odds_M}{Odds_F} = 3{,}255$, tregon se meshkujt kanë 3,255 herë shanse më shumë për të mos qënë të përfshirë në skemën e Sigurimeve Shoqërore në krahasim me femrat. Vlera e $\ln(3{,}255) = 1{,}1801$ shërben si koeficient i variablit të pavarur $Gjinia$ tek ekuacionit i regresit logjistik të thjeshtë.
Trajta e ekuacionit:

$$logit(P) = -1.699 + 1{,}1801 \cdot Gjinia$$

Gjithashtu për të vlerësuar lidhjen midis përfshirjes në skemën shoqërore dhe $Gjinis$ë u vlerësohet modeli i regresit logjistik:

$$Logit(\pi) = Ln(Odds) = \beta_0 + \beta_1 \cdot Gjinia$$

Përfundimet e marra nga zbatimi në SPSS janë:

Tabela 4.7: Testi Omnibus për koeficientët e modelit
Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Hapi 1	Step	185.641	1	.000
	Block	185.641	1	.000
	Model	185.641	1	.000

Vemë re

$$-2 \cdot LL = -2\ln(raporti\ i\ pergjasise) = -2\left(\frac{L_0}{L_1}\right) = 185.641$$

Hipotezat që formulojmë për testin Omnibus janë:
H_0: modeli nuk është i rëndësishëm
H_a: modeli është i rëndësishëm

Nga testi Omnibus vihet re se modeli i ndertuar është statistikisht i rëndësishëm me nivel rëndësie $\alpha = 0{,}05$.

Tabela 4.8: Treguesit e modelit te regresit logjistik

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3834.274[a]	.055	.078

Vlerat e Cox & Snell R Square R^2 dhe Nagelkerke R^2 janë të vogla edhe pse modeli rezulton që është statistikisht i rëndësishme dhe kjo vjen si arsye e mungesës së variableve të tjerë në model të cilët mund të kenë ndikim më të madh shpjegues.

Tabela 4.9: Tabela e koeficientëve të ndryshoreve në ekuacion

Variablat në ekuacion

		B	S.E.	Wald	df	Sig.	Exp(B)	95% Conf. Interval	
Step 1[a]	Gjinia	1.180	0.092	163.203	1	.000	3.255	0.999718	1.360282445
	Constant	-1.699	0.081	441.825	1	.000	0.183	-1.85773	-1.540273065

a. Variable(s) entered on step 1: gjinia.

Nga tabela shohim se të dy variablat janë statistikisht të rëndësishëm dhe merret ky model i regresit logjistik:

$$Logit(\pi(x)) = -1{,}699 + 1{,}180 \cdot Gjinia$$

Meqenëse $\beta_1$tregon se shanset që një mashkull të mos jetë i përfshirë në skemë janë më të larta se të një femre, β_1 është ndryshimi mesatar në probabilitetin e mos përfshirjes në skemë të një individi midis Meshkujve dhe Femrave

4.4.2 Rasti i dytë (Varësia e ndryshores kategorike me një ndryshore të vazhdueshme)

Në këtë rast është vrojtuar varësia e ndryshores sonë kategorike nga ndryshorja e vazhdueshme mosha e të intervistuarit.

Tabela 4.10: Testi Omnibus për koeficientët e modelit

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	66.026	1	.000
	Block	66.026	1	.000
	Model	66.026	1	.000

Vemë re

$$-2 \cdot LL = -2\ln(raporti\ i\ pergjasise) = -2\left(\frac{L_0}{L_1}\right) = 66.026$$

Hipotezat që formulojmë për testin Omnibus janë:
H_0: modeli nuk është i rëndësishëm
H_a: modeli është i rëndësishëm
Nga testi Omnibus vihet re se modeli i ndertuar është statistikisht i rëndësishëm.

Tabela 4.11: Treguesit e modelit te regresit logjistik

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3953.889[a]	.020	.028

a

Përsëri vlerat e Cox & Snell R Square R^2 dhe Nagelkerke R^2 janë të vogla edhe pse modeli rezulton që është statistikisht i rëndësishme dhe kjo vjen si arsye e mungesës së variableve të tjerë në model të cilët mund të kenë ndikim më të madh shpjegues.

Tabela 4.12: Tabela e koeficientëve të ndryshoreve në ekuacion

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1[a]	Mosha	-.026	.003	64.745	1	.000	.974
	Constant	.226	.139	2.649	1	.104	1.254

a. Variable(s) entered on step 1: mosha.

Modeli i regresit logjistik të thjeshtë është:

$$Logit\big(\pi(x)\big) = -0{,}026 + 0{,}226 \cdot Mosha$$

Nga tabela vemë re se me rritjen e moshës me një vit shanset që një individ të mos jete i përfshirë në skemë ulen mesatarisht me 0.974.

4.4.3 Rasti i tretë (Varësia e ndryshores kategorike me ndryshore të vazhdueshme, kategorike)

Nga studimi me kujdes i bazës së të dhënave u gjetën disa ndryshore të tjera të rëndësihme sic janë gjinia, mosha, sektori i punësimit, vitet e shkollimit dhe zona.
Tre ndryshore janë kategorike dhe madje me disa opsione SPSS i kodon duke i futur në model si variabla të rinj por të koduar me 0, 1. Domethënë nga 6 kategori te ndryshorja lloji i punësimit në model do të shtohen 5 ndryshore, nga 4 kategori tek ndryshorja zona do të shtohen 3 ndryshore. Në tabelën e mëposhtme jepen këto kodime për secilën nga ndryshoret e marra në studim. Mungesa e informacionit në disa nga variablat e rinj të shtuar në model ka reduktuar numrin e individeve të cilët janë marrë në shqytim. Në total numri i individëve për modelin e regresit logjistik është 3166.

Tabela 4.13: Kodimi i ndryshoreve kategorike

Kodimi i ndryshoreve kategoreike							
		Dendutirë	Kodimi i parametrave				
			(1)	(2)	(3)	(4)	(5)
Lloji i punësimi	Sektori publik	1235	1.000	.000	.000	.000	.000
	Kompani private	1429	.000	1.000	.000	.000	.000
	Program i punës publike	19	.000	.000	1.000	.000	.000
	Ndërmarrje shtetërore	76	.000	.000	.000	1.000	.000
	OJQ ose Organizatë humanitare	16	.000	.000	.000	.000	1.000
	Vetëpunësuar	391	.000	.000	.000	.000	.000
Zona	Qëndrore	1226	1.000	.000	.000		
	Bregdetare	961	.000	1.000	.000		

	Malore	402	.000	.000	1.000		
	Tirana	577	.000	.000	.000		
Gjinia	Male	2042	1.000				
	Female	1124	.000				

Rezultatet e perftuara nga SPSS për këtë rast janë:

Tabela 4.14: Testi Omnibus për koeficientët e modelit

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1513.295	11	.000
	Block	1513.295	11	.000
	Model	1513.295	11	.000

Nga tabela vemë re se modeli i ndërtuar është statistikisht i rëndësishëm (sig=0.000<0.05)

Tabela 4.15: Treguesit e modelit të regresit logjistik

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2277.563[a]	.380	.544

Vihet re se me shtimin e ndryshoreve të reja në model rriten në mënyrë të ndjeshme vlerat e Cox & Snell R^2 dhe Nagelkerke R^2.
Vlerat e dy koeficientëve R^2 tregojnë se modeli shpjegon rreth 38% deri në 54.4% të shpërhapjes së të dhënave.

Tabela 4.16: Tabela e klasifikimit

	Vrojtimi		Parashikimi		
			Skema e Sigurimeve Shoqërore		Përqindjet korrekt
			Po	Jo	
Step 1	Skema e Sigurimeve Shoqërore	Po	2059	201	91.1
		Jo	356	550	60.7
	Përqindja totale e klasifikimit korrekt				82.4

Gabimi të llojit të parë kemi kur një individ është parashikuar nga modeli që nuk përfshihet në skemë ndërkohë që ai realisht është i përfshirë në skemë.
Konkretisht probabiliteti i gabimit llojit të parë është 8.9% (201/(2059+201)). Gabim të llojit të dytë kemi kur një individ është parashikuar nga modeli si i përfshirë në skemë ndërkohë që ai realisht nuk është i përfshirë në skemë 39.3%(356/(356+550)).
Përqindjet e sakta për ato të cilët janë të përfshirë ose nuk janë të përfshurë janë përkatësisht 91.1%, 60.7% . Pra modeli në rastet kur një individ realisht është i përfshirë në skemë arrin të parashikoj saktë 91.1% të rasteve, ndëkohë kur një individ nuk është i përshirë në skemë modeli arrin të parashikoj saktë në 60.7% të rasteve. Përqindaja totale e parashikimi të saktë është një mesatare e peshuar midis dy përqindjeve të sakta.

Tabela 4.17: Tabela e koeficientëve të ndryshoreve në ekuacion
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gjinia	.793	.122	41.988	1	.000	2.211
	Mosha	-.021	.005	21.669	1	.000	.979
	Lloji i punësimit			417.176	5	.000	
	Sektori publik	-4.835	.244	391.445	1	.000	.008
	Kompani private	-2.381	.168	200.023	1	.000	.092

	Program i punës publike	-3.429	.807	18.044	1	.000	.032
	Ndërmarrje shtetërore	-4.632	.743	38.878	1	.000	.010
	OJQ ose organizata humanitare	-1.213	.544	4.967	1	.026	.297
	Vite_shkollimi	-.204	.017	138.525	1	.000	.816
	Zona			9.101	3	.028	
	Qëndrore	-.315	.152	4.315	1	.038	.730
	Bregdetare	.018	.155	.014	1	.905	1.019
	Malore	.020	.212	.009	1	.926	1.020
	Constant	4.574	.373	150.259	1	.000	96.969

a. Variable(s) entered on step 1: Gjinia, Mosha, Lloji i punësimi, vitet e shkollimit, zona.

Nga rezultatet e përftuara vihet re se në përjashtim të ndryshores Zona (2) dhe zona (3) të gjithë të ndryshoret e tjera janë statistikisht të rëndësishme (sig<0.05) dhe merret modeli i regresit të shumëfishtë logjistik:

$Logit(\pi(x))$ = 4.574+0.793**gjinia* -0.021**mosha* -4.835* *sektori publik* -2.381* *kompani private* -3.429* *programi punës publike* -4.632**ndërmarrje shtetërore*−1.213* *OJQ* -0.204 *vite shkollimi* - 0.315**zona qëndrore* +0.018* *zona bregdetare* +0.020* *zona malore* (4.1)

Për një mashkull (mashkull=1), mosha =30 vjec, punësimi në kompani private, 16 vite shkollim dhe zona bregdetare logit i parashikuar është:

$Logit\{$ mashkull=1, 30, punësimi në kompani private = 1, 16, zona(bregdetare)=1}= -0.8899

Probabiliteti që një mashkull me këto karakteristika të mos përfshihet në skemën e sigurimeve Shoqërore është:

$$P(y=1)=\frac{e^{-0.8899}}{1+e^{-0.8899}}=0.291$$

Le të bëjmë intepretimin e koeficientëve pranë ndryshoreve tek modeli i regresit logjistik

Trajta e përgjithshme e modelit të ndërtuar të regresit është:

$$logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53}$$

- Interpretimi i koeficientit pranë moshës si rast i një ndryshoreje të vazhdueshme.

Intepretimi i koeficientit β_2 tek modeli i regresit logjistik të shumëfishtë bëhet në këtë mënyrë:

Shënojmë me k një vlerë e cfarëdoshme të x_2 :

$x_2 = k$, fiksojmë vlerat e $x_1, x_{31}, \dots, x_{53}$,

$$logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 k + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53}$$

$$Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 k + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53}}$$

$x_2 = k + 1$,

$$logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 (k+1) + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53}$$

$$Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 (k+1) + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53}}$$

$$Odds = \frac{Odds\ kur\ x_2 = k+1}{Odds\ kur\ x_2 = k}$$

$$= \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 (k+1) + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53}}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 k + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53}}}$$

$$= e^{\beta_2}$$

Me rritjen një vit të moshës dhe mbajtjen konstante të ndryshoreve të tjera shanset që një individ të mos jëtë i përfshirë në skemë zvogëlohen më 0.979 herë.

- Interpretimi i ndryshores lloji i punësimit

Interpretimi i koeficientëve kur ndyshorja është kategorike me disa opsione:

Kur ndryshorja është kategorike duhet të kemi gjithmonë parasysh kodimin e ofruar nga programi. Konkretisht nëse punonjësi është në sektorin publik ndryshorja e formuar nga kjo

kategori do marr vlerën 1 dhe gjithë ndryshoret e tjera të përftuara nga kategoritë e punësimit do të marrin vlerën zero. Prandaj do të interpretohet vetëm koeficienti pranë punësimit në kompani private. Kështu që shanset që një punonjës i punësuar në kompani private krahasuar me një punojës të vetëpunësuar të mos jetë i përfshirë në skemë zvogëlohen 0.008 herë.

$$logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{31} + \beta_4 x_{32} + \beta_5 x_{33} + \beta_6 x_{34} + \beta_7 x_{35} + \beta_8 x_4 + \beta_9 x_{51} + \beta_{10} x_{52} + \beta_{11} x_{53} \quad (4.2)$$

Nryshorja sektori publik nga kodimi ka vlerën $x_{31} = 1$ ndërsa ndryshoret e tjera $x_{32} = x_{33} = x_{34} = x_{35} = 0$

Ndryshorja kompani private nga kodimi ka vlerën $x_{32} = 1$ ndërsa ndryshoret e tjera $x_{31} = x_{33} = x_{34} = x_{35} = 0$.

Nga ekuacioni (4.2) marrim:

$$logit(\pi(x)) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot 1 + \beta_4 \cdot 0 + \beta_5 \cdot 0 + \beta_6 \cdot 0 + \beta_7 \cdot 0 + \beta_8 \cdot x_4 + \beta_9 \cdot x_{51} + \beta_{10} \cdot x_{52} + \beta_{11} \cdot x_{53}$$

dhe

$$logit(\pi(x)) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot 0 + \beta_4 \cdot 1 + \beta_5 \cdot 0 + \beta_6 \cdot 0 + \beta_7 \cdot 0 + \beta_8 \cdot x_4 + \beta_9 \cdot x_{51} + \beta_{10} \cdot x_{52} + \beta_{11} \cdot x_{53}$$

$$Odds = \frac{e^{\beta_0+\beta_1\cdot x_1+\beta_2\cdot x_2+\beta_3\cdot 1+\beta_4\cdot 0+\beta_5\cdot 0+\beta_6\cdot 0+\beta_7\cdot 0+\beta_8\cdot x_4+\beta_9\cdot x_{51}+\beta_{10}\cdot x_{52}+\beta_{11}\cdot x_{53}}}{e^{\beta_0+\beta_1\cdot x_1+\beta_2\cdot x_2+\beta_3\cdot 0+\beta_4\cdot 1+\beta_5\cdot 0+\beta_6\cdot 0+\beta_7\cdot 0+\beta_8\cdot x_4+\beta_9\cdot x_{51}+\beta_{10}\cdot x_{52}+\beta_{11}\cdot x_{53}}} = \frac{e^{\beta_3}}{e^{\beta_4}}$$

$$= e^{-4.835-(-2.381)}$$

$$= e^{-2.454} = 0.0859$$

Shanset që individi i punësuar në sektorin publik të mos jetë i përfshirë në Skemën e Sigurimeve Shoqërore janë 0.0859 herë më të ulta se shanset që një individ i punësuar në kompani private.

4.5 Përfundime

Në këtë kapitull realizohet modelimi i të dhënave me regresin logjistik të shumëfishtë. Të dhënat e përdorura në modelim janë marrë nga baza e të dhënave të The World Bank (2012). Ato përdoren për të parashikuar probabilitetin e mos përfshirjes së një individi të punësuar në skemën e sigurimeve shoqërore. Ndërtohen dhe analizohen modele të regresit të thjeshtë dhe regresit të shumëfishtë logjistik duke konsideruar si ndryshore të pavarura: gjinia, mosha, lloji i punësimit, vitet e shkollimit dhe zona. Ndryshoret lloji i punësimit dhe zona u koduan si ndryshore kategorike me 0 dhe 1 dhe futën në model si ndryshore të reja. Kështu nga 6 kategori te ndryshorja lloji i punësimit në model shtohen 5 ndryshore të reja ndërsa për zonën shtohen 3 ndryshore. Pra si ndryshore të pavarur në modelin e regresit të shumëfishtë logjistik janë: gjinia, mosha, sektori publik, kompani private, program i punës publike, ndërmarrje shtetërore, OJQ ose organizata humanitare, vitet e shkollimit, zona qëndrore, bregdetare, malore.

Nga trajtimi i modeleve të ndërtuara të regresit logjistik të thjeshtë dhe shumëfishtë përcaktohet modeli statistikisht më i mirë i cili parashikon probabilitetin e mos përfshirjes së një individi të punësuar në skemën e sigurimeve shoqërore. Modeli wshtw:

$Logit(\pi(x))$ = 4.574+0.793**gjinia* -0.021**mosha* -4.835* *sektori publik* -2.381* *kompani private* -3.429* *programi punës publike* -4.632**ndërmarrje shtetërore*−1.213* *OJQ* -0.204 *vite shkollimi* - 0.315**zona qëndrore* +0.018* *zona bregdetare* +0.020* *zona malore*

Duke analizuar modelet e regresit logjistik të shumëfishtë arrijmë në përfundimin se me shtimin e ndryshoreve të pavarura në model rritet përqindja totale e klasifikimit korrekt. Në modelin e regresit të shumëfishtë logjistik u përcaktuan si variabla satistikisht të rëndësishëm: gjinia, mosha, sektori publik, kompani private, program i punës publike, ndërmarrje shtetërore, OJQ ose organizata humanitare, vitet e shkollimit, zona qëndrore, bregdetare. Modeli i ndërtuar shpjegon 38% deri në 54.4% të shpërhapjes së të dhënave dhe përqindja totale e klasifikimit korrekt është 82.8%.

REFERENCAT

[1] Agresti, A Categorical Data Analysis, 2nd ed., University of Florida, 2002, pp.1-36, 165-196.

[2] Agresti, A and Hitchcock, D. Statistical Methods & Applications (2005) 14: 297–330DOI: 10.1007/s10260-005-0121-y,"Bayesian inference for categorical data analysis".

[3] Agresti, A. and Chuang, C. (1989) Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. Computational Statistics and Data Analysis, 7, 245-258.

[4] Agresti, A. and Min, Y. (2005) Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 x2 contingency tables. Biometrics, 61, 515-523.

[5] Agresti, A. (1989). A survey of models for repeated ordered categorical response data, Statistics in Medicine, 8, 1209-1224.

[6] Agresti, A (1992). A survey of exact inference for contingency tables, Statistical Science, 7(1), 131-177.

[7] Aitchison, J. (1985) Practical Bayesian problems in simplex sample spaces. In Bayesian Statistics 2, 15-31. North-Holland/Elsevier (Amsterdam; Neë York).

[8] Albert, J. H. (1984) Empirical Bayes estimation of a set of binomial probabilities. Journal of Statistical Computation and Simulation, 20, 129-144.

[9] Bernardo, J. M. and Ram_on, J. M. (1998) An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. The Statistician, 47, 101-135.

[10] Cenaj, E, Prifti, L, The role of statistical independence in contingency tables, Vol.3, ISSUE 10, October 2016. ISSN: 2458-9403

[11] Cenaj, E, Dervishi,R, Muka, Zh **"**The use of logistic regression in a human health problem"IJISET, International Journal of Innovative Science, Engineering& Technology, Vol.3 Issue 12, Decembre 2016, ISSN (Online) 2348-7968, ëëë.ijiset.com

[12] Cenaj, E, Prifti, L, "Statistical Analysis of effect of air pollution on human health" 7th International Conference of Ecosystems, June 01-03 2017, Tirana, Albania.

[13] Ceanj, E, Dervishi, R, "Using Bayesian Methods for Categorical Data Analysis" ICASE 2017, International Conference on Applied satistics and Econometrics, 27-28 April 2017, Tirana, Albania.

[14] Delampady, M. and Berger, J. O. (1990) Lower bounds on Bayes factors for multinomial distributions, ëith application to chi-squared tests of _t. The Annals of Statistics, 18, 1295-1316.

[15] Diaconis, P. and Freedman, D. (1990) On the uniform consistency of Bayes estimates for multinomial probabilities. The Annals of Statistics, 18, 1317-1327.

[16] Ericson, W. A. (1969) Subjective Bayesian models in sampling finite populations. Journal of the Royal Statistical Society, Series B, Methodological, 31, 195-233.

[17] Good, (1965) The Estimation of Probabilities: An Essay on Modern Bayesian Methods. Cambridge, MA: MIT Press.

[18] Good, I. J. and Crook, J. F. (1974) The Bayes/non-Bayes compromise and the multinomial distribution. Journal of the American Statistical Association, 69, 711-720.

[19] Hosmer.D.V, Lemeshow.S, (2000). Applied Logistic Regression. Second Edition, John Wiley & Sons. pp 1-45.

[20] Hoadley, B. (1969) The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. Journal of the American Statistical Association, 64, 216-229.

[21] Insituti i Sigurimeve Shoqërore, Statistika të Sigurimeve Shoqërore

[22] Johnson, B. M. (1971) On the admissible estimators for certain fixed sample binomial problems. The Annals of Mathematical Statistics, 42, 1579-1587.

[23] Kamberi. L, Tabelat e kontigjencës, Zbatime. 2012

[24] Leka. Sh, "Teoria e Probabiliteteve dhe Statistika matematike" SHBLU, 2004.

[25] Lindley, D. V. (1964) The Bayesian analysis of contingency tables. The Annals of Mathematical Statistics, 35, 1622-1643.

[26] Madigan, D. and York, J. C. (1997) Bayesian methods for estimation of the size of a closed population. Biometrika, 84, 19-31.

[27] Naqo, M, "Statistika Matematike" SHBLU 2005

[28] Perks, W. (1947) Some observations on inverse probability including a new indi_erence rule. Journal of the Institute of Actuaries, 73, 285-334.

[29] Prifti, L, Shehu, Sh, "Modeling categorical data in a human health problem" NTAM, 7-8 November 2014.
[30] Open Data, Albania
[31] Routledge, R. D. (1994) Practicing safe statistics with the mid-p*. The Canadian Journal of Statistics, 22, 103-110.

[32] Rukhin, A. L. (1988) Estimating the loss of estimators of a binomial parameter. Biometrika, 75, 153-155.

[33] Smith, P. J. (1991) Bayesian analyses for a multiple capture-recapture model. Biometrika, 78, 399-407.

[34] Stakes. E. Maura, Davis. S. Charles, Koch. G. Gary, categorical data analysis using the SAS system, 2nd Edition.

[35] Stephen, E, Fienberg "When Did Bayesian Inference Become "Bayesian" "2006 Internatinoal Society for Bayesian Analysis.

[36] Simon P. Washington, Matthew G. Karlaftis, Fred Mannering Statistical and Econometric Methods for Transportation Data Analysis, Second Edition, 2010.

[37] Salillari, D. "Modelimi matematik i gërmave në tekstet shqip dhe zbatime në sisteme kompjuterike" 2016.

[38] Sh. Tsumoto, Statistical Independence as Linear Independence Published by Science B. V, March 2003, Volume 82, Issue 4,vol. 82., pp.274-285.

[39] Sh. Tsumoto. Statistical independence and Determinants in a Contingency table.Interpretation of Pearson Reziduals based on linear Algebra. Volume 90, Issue 3, March 2009 pp251-267.

[40] Sh. Tsumoto, S. Hirano, Role of Marginal Distribution in a Contingency table, Proceedings of NAFIPS 2006, IEEE press, 2006.

[41] Sh. Tsumoto, Sh. Hirano, Decomposition of Contingency Table as Tensor Product, Published in 2006 IEEE International Conference on Fuzzy Systems, 16-21 July 2006.

[42] The World Bank, Albania - Living Standards Measurement Survey 2012