VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

Vaida Raščiauskaitė

Pirkinių krepšelio statistiniai modeliai

Magistro baigiamasis darbas

VILNIUS, 2012

Darbo vadovas: _____
Doc. Vytautas Kazakevičius (parašas)

Darbo recenzentas: _____
Doc. Rimantas Eidukevičius (parašas)

Registracijos Nr.: .....
Darbo gynimo data: 2012 birželio 4 d.

Turinys

Įvadas

Šiais laikais įmonės nuolat ieško būdų, kaip padaryti pardavimus efektyvesniais. Įmonei, parduodančiai prekes telefonu, tai reiškia parduoti kuo daugiau prekių su kuo mažesnėmis išlaidomis telefonų pokalbiams. Taip pat svarbu atrinkti siūlomas prekes, nes užvertus klientą pasiūlymais, kurie jo visiškai nedomina, galima jį nesunkiai prarasti.

Norint duoti moksliškai pagrįstas rekomendacijas vienai ar kitai įmonei, kuriami vadinamojo *pirkinių krepšelio* statistiniai modeliai. Savo magistro darbe aš nagrinėju keletą tokių modelių, aiškinuosi, kaip vertinami jų parametrai naudojant jau egzistuojančią programinę įrangą, ir pritaikau teoriją modeliuodama vienos realios įmonės pardavimų procesą.

Darbo struktūra tokia: 1 skyriuje aprašomi turimi duomenis ir formuluojami uždaviniai. 2 skyriuje pabandau įvertinti krepšelių parametrus ir analizuoju trys pirkinių krešelio modeliai. Pirmajame laikoma, kad prekės į krepšelį dedamos atsitiktinai. Antrasis yra apibendrintos logistinės regresijos modelis, paimtas iš literatūros, kuriame ieškoma ryšių tarp prekių. Trečiasis – pačios sukurtas logistinės regresijos modelis, atsižvelgiantis į kliento norą pirkti prekę ir prekės kainą.

1 Duomenys

1.1 Duomenų lentelės struktūra

Įmonė, kurios duomenis tiriame, pardavinėja prekes telefonu. Dažniausiai siūloma pirkti knygas, proginius medalius arba monetas, apatinį trikotažą, maisto papildus. Vieno skambučio metu siūloma viena prekė. Užsakytos prekės yra siunčiamos klientui per kurjerį arba į pašto skyrių. Įmonės duomenų bazėje saugoma tokia informacija:

- kliento kodas;
- informacija apie klientą;
- skambučio data;
- siūlytos prekės kodas;
- prekės kaina;
- užsakymo data;
- apmokėjimo data;
- grąžinimo data.

Tyrimui buvo pasirinkti duomenys apie 4 knygų pardavimus nuo 2010 metų liepos ir 2012 metų balandžio. Duomenų lentelėje yra 10365 eilutės ir 6 stulpeliai:

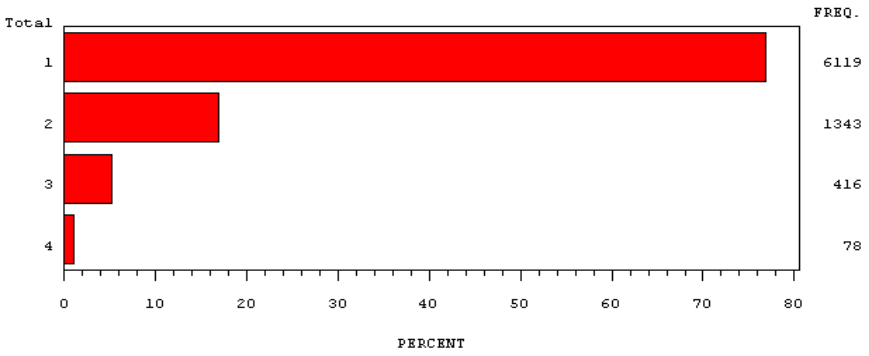
- kliento kodas;
- laiko intervalas nuo paskutinio skambučio;
- siūlytos prekės kodas;
- prekės kaina;
- ar prekė užsakyta;
- laiko intervalas nuo paskutinio mokėjimo.

Viena eilutė atitinka vieną skambutį klientui.

1.2 Pagrindinis duomenų charakteristikos

1.2.1 Skambučiai

Nagrinėjamu laikotarpiu buvo skambinta 7956 unikaliems klientams, kiekvienam – nuo 1 iki 4 kartų, žr. 1 pav.



1 pav.: Skambučių skaičiaus 1 klientui pasiskirstymas

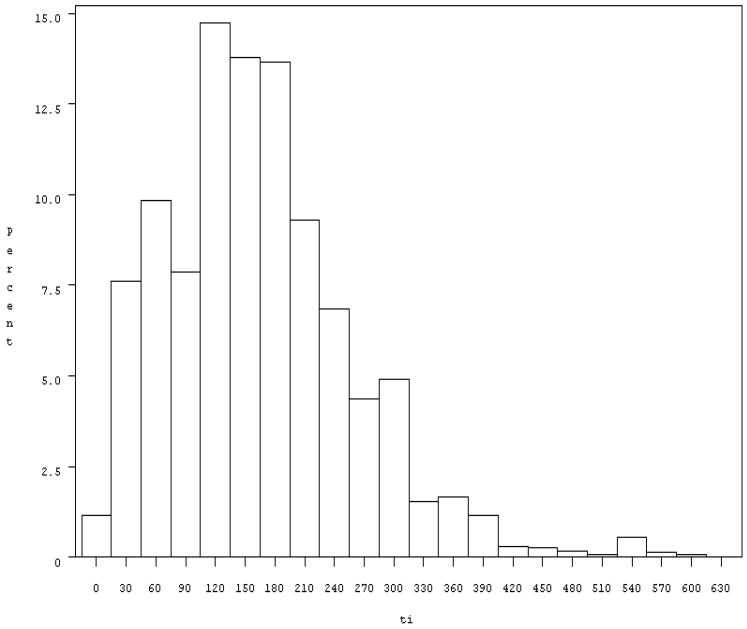
2 pav. parodyta, kaip pasiskirstę intervalai tarp skambučių vienam klientui. Intervalo vidurkis yra 163.7 dienos (95% pasikliautinas intervalas: 160–167.5), ilgiausias intervalas buvo 602 dienos (apytiksliai metai ir 8 mėnesiai), trumpiausias – 1 diena.

3 pav. matome analogišką intervalų tarp apmokėjimų pasiskirstymą. Jei klientas pageidauja prekę atsiimti pašte, apmokėjimas už ją gali užtrukti, nes paštas saugo siuntinius iki 2 savaičių. Todėl intervalai tarp prekių užsakymų ir apmokėjimų gali skirtis. Intervalo tarp apmokėjimų vidurkis yra 144.6 dienos (95% pasikliautinas intervalas: 139.6–149.7 dienos), ilgiausias tarpas tarp apmokėjimų buvo 622 dienos, trumpiausias – 1 diena.

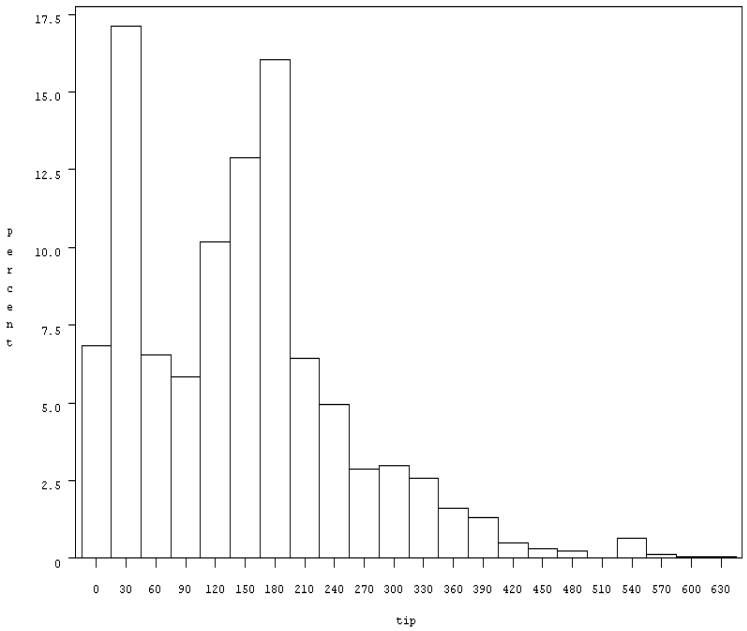
1.2.2 Prekės ir pirkinių krepšeliai

Kaip minėjau 1.1 skyrelyje, darbe nagrinėjau 4 knygų pardavimus. Toliau knygas numeruosiu skaičiais nuo 1 iki 4. 1 lentelėje surašytos pagrindinės jų charakteristikos.

4 pav. parodyta, kiek klientų pirko arba nepirko pasiūlytą knygą, o 5 pav. – kiek knygų yra nusipirkę tiriami klientai. Tikslesnę informaciją apie pirkinių krepšelius duoda 2 lentelė.



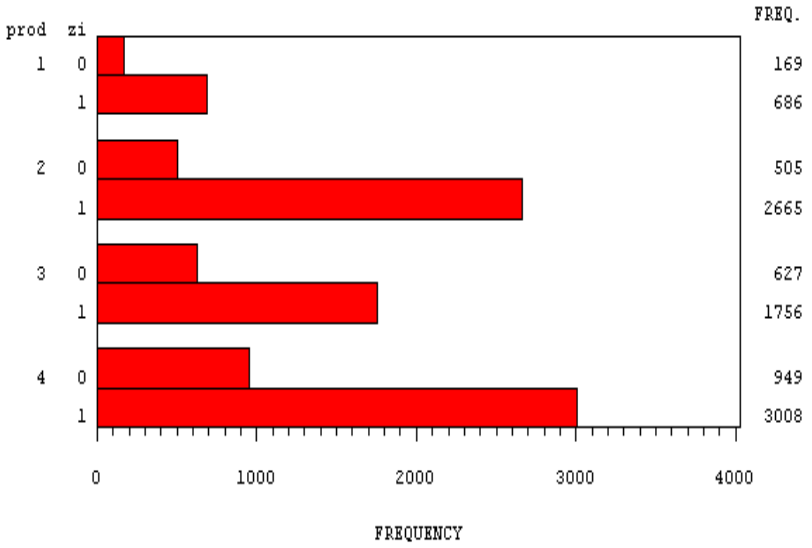
2 pav.: Intervalų tarp skambučių vienam klientui pasiskirstymas



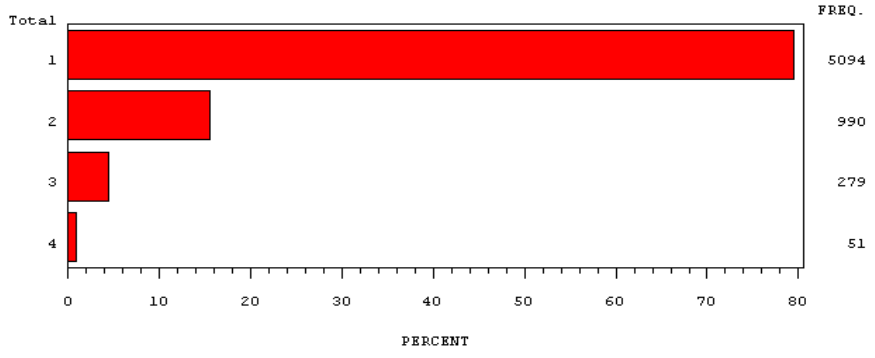
3 pav.: Intervalų tarp apmokėjimų pasiskirstymas

Knygos kodas	Tematika	Kaina (Lt)
1	apie augalus	87
2	apie konservavimą	83
3	apie gydymąsi	79
4	apie sveiką gyvenseną	54

1 lentelė: Knygos



4 pav.: Kiek klientų pirko pasiūlytą knygą



5 pav.: Nupirktų knygų skaičiaus pasiskirstymas

Krepšelis	Krepšelių skaičius	Krepšelių dalis (proc.)
$\emptyset$	1542	19,38
$\{1\}$	299	3,76
$\{2\}$	1781	22,38
$\{3\}$	960	12,06
$\{4\}$	2055	25,83
$\{1,2\}$	86	1,08
$\{1,3\}$	66	0,83
$\{1,4\}$	64	0,8
$\{2,3\}$	174	2,19
$\{2,4\}$	326	4,1
$\{3,4\}$	274	3,44
$\{1,2,3\}$	40	0,5
$\{1,2,4\}$	48	0,6
$\{1,3,4\}$	32	0,4
$\{1,2,3,4\}$	51	0,64

2 lentelė: Pirkinių krepšelio skirstinys

2 Pirkinių krepšelio modeliai

2.1 Pagrindinės sąvokos

Tegu I žymi *prekių aibę*; jos elementus (prekes) žymėsiu i ir j raidėmis. Apskritai I gali būti bet kokia aibė, tačiau dažniausiai laikoma, kad jos elementai yra pirmi N natūralieji skaičiai:

$$I = \{1, 2, \ldots, N\}.$$

Pirkinių krepšeliu (toliau – tiesiog *krepšeliu*) vadinamas bet koks I aibės poaibis. *Krepšeliai* paprastai žymimi b raidėmis. Pavyzdžiui,

$$b_1 = \emptyset, \quad b_2 = \{1, 3\}$$

ir pan.

Tegu stebima n pirkėjų ir b_k žymi k-tojo pirkėjo krepšelį. Literatūroje (žr., pavyzdžiui, [1]) krepšeliai $b_1, \ldots, b_n$ paprastai laikomi tam tikro atsitiktinio krepšelio B nepriklausomomis realizacijomis. B krepšelio skirstinys apibrėžiamas, nurodant kiekvienos galimos reikšmės b tikimybę

$$p(b) = \mathrm{P}\{B = b\}.$$

Kadangi iš viso yra 2^N galimų krepšelių b, apibrėžiant skirstinį reikia nurodyti $2^N - 1$ tikimybę. Taigi, gaunamas standartinis statistinis uždavinys: turint paprastąją imtį iš B skirstinio, reikia įvertinti $2^N - 1$ skirstinio parametrą.

Su bet kokiu krepšeliu b ir bet kokiu $i \in I$ apibrėžiu

$$b(i) = \begin{cases} 1, & \text{kai } i \in b; \\ 0, & \text{kai } i \notin b. \end{cases}$$

Jei žinomas b, žinomas ir skaičių rinkinys $\Big(b(1), \ldots, b(N)\Big)$. Atvirkščiai, tas skaičių rinkinys visiškai apibrėžia krepšelį b:

$$b = \{i \mid b(i) = 1\}.$$

Jei B yra atsitiktinis krepšelis, tai $\Big(B(1), \ldots, B(N)\Big)$ yra atsitiktinis vektorius; jo koordinatės $B(i)$ yra atsitiktiniai dydžiai, įgyjantys dvi reikšmes 0 ir 1. Vektoriaus skirstinys aprašomas, nurodant tikimybes

$$p(u_1, \ldots, u_N) = P\{B(1) = u_1, \ldots, B(N) = u_N\};$$

čia $u_1, \ldots, u_N \in \{0, 1\}$. Taigi atsitiktinis krepšelis faktiškai yra N-matis atsitiktinis vektorius, kurio koordinatės įgyja dvi reikšmes 0 ir 1.

Mano nagrinėjamoje situacijoje $N = 4$ ir

$$I = \{1, 2, 3, 4\}.$$

Galėčiau laikyti, kad b_k yra k-ojo pirkėjo nupirktų knygų aibė. Bet tada prielaida, kad $b_1, \ldots, b_n$ yra paprastoji imtis iš B skirstinio, nėra korektiška, nes ne visiems klientams buvo pasiūlytos visos knygos. Todėl norint pilnai aprašyti turimus duomenis reikia pateikti informaciją ir apie nesiūlytas prekes.

Nagrinėkime dvi prekes 1 ir 2. Tarkime, su $s = 1, 2$ atsitiktinis dydis $Y_k(s)$, kuris įgyja reikšmę 1, jei klientui buvo pasiūlyta prekė, ir 0, jei nebuvo pasiūlyta, o atsitiktinis dydis $X_k(s)$ įgyja reikšmę 1, jei klientas pirko prekę, ir 0, jei nepirko. Galime laikyti, kad visi $(Y_k(1), Y_k(2), X_k(1), X_k(2))$, $k = 1, \ldots, n$, vektoriai yra nepriklausomi ir vienodai pasiskirstę. Tegu $Y_k = (Y_k(1), Y_k(2))$ skirstinys aprašomas lentele

Y_k	Tikimybės
(1,0)	q_{10}
(0,1)	q_{01}
(1,1)	$q_{11} = 1 - q_{10} - q_{01}$

o sąlyginis $X_k = (X_k(1), X_k(2))$ skirstinys Y_k atžvilgiu – lentele

Y_k	X_k	Tikimybės
(1,0)	(1,0)	$p_{1.}$
	(0,0)	$p_{0.}$
(0,1)	(0,1)	$p_{.1}$
	(0,0)	$p_{.0}$
(1,1)	(0,0)	p_{00}
	(0,1)	p_{01}
	(1,0)	p_{10}
	(1,1)	p_{11}

2.2 Krepšelių parametrų vertinimas

2.2.1 Klasikinė situacija

Tarkime, turime n dydžio krepšelių imtį $b_1, ..., b_n$ iš skirstinio B su parametrais $p(b)$. Gerai žinome, kad tada parametrų didžiausio tikėtinumo įverčiai skaičiuojami pagal formulę

$$\hat{p}(b) = \frac{n(b)}{n};$$

čia $n(b)$ žymi b krepšelio pasikartojimų skaičių imtyje.

2.2.2 Netipinių krepšelių situacija

Jei ne visiems klientams buvo siūlytos abi prekės, tariame, kad mūsų duomenys yra (Y_k, X_k) vektorių rinkinys, kuris yra paprastoji imtis iš skirstinio aprašyto 2.1 skyrelyje. Duomenys gali būti sutraukti į tokią dažnių lentelę:

(Y_k, X_k)	Krepšelių sk.	Tikimybės
(1,0,0,0)	$n_{0.}$	$q_{10}(1-p_{1.})$
(1,0,1,0)	$n_{1.}$	$q_{10}p_{1.}$
(0,1,0,0)	$n_{.0}$	$q_{01}(1-p_{.1})$
(0,1,0,1)	$n_{.1}$	$q_{01}p_{.1}$
(1,1,0,0)	n_{00}	$(1-q_{10}-q_{01})(1-p_{10}-p_{01}-p_{11})$
(1,1,1,0)	n_{10}	$(1-q_{10}-q_{01})p_{10}$
(1,1,0,1)	n_{01}	$(1-q_{10}-q_{01})p_{01}$
(1,1,1,1)	n_{11}	$(1-q_{10}-q_{01})p_{11}$

Tikėtinumo funkcija

$$L = q_{10}^{n_{0.}+n_{1.}} q_{01}^{n_{.0}+n_{.1}} (1-q_{10}-q_{01})^{n_{++}} (1-p_{1.})^{n_{0.}} p_{1.}^{n_{1.}} \times \\ \times (1-p_{.1})^{n_{.0}} p_{.1}^{n_{.1}} (1-p_{10}-p_{01}-p_{11})^{n_{00}} p_{10}^{n_{10}} p_{01}^{n_{01}} p_{11}^{n_{11}},$$

o jos logaritmas

$$\begin{aligned}\ell = n_{+.}\ln q_{10} + n_{.+}\ln q_{01} + n_{++}\ln(1-q_{10}-q_{01}) + \\ n_{0.}\ln(1-p_{1.}) + n_{1.}\ln p_{1.} + n_{.0}\ln(1-p_{.1}) + n_{.1}\ln p_{.1} + \\ n_{00}\ln(1-p_{10}-p_{01}-p_{11}) + n_{10}\ln p_{10} + n_{01}\ln p_{01} + n_{11}\ln p_{11};\end{aligned}$$

čia

$$n_{+.} = n_{0.} + n_{1.},$$

$$n_{.+} = n_{.0} + n_{.1},$$

$$n_{++} = n_{00} + n_{01} + n_{10} + n_{11}.$$

Suskaičiuojame dalines ℓ funkcijos išvestines ir prilyginę jas 0, gauname tokią lygčių sistemą:

$$\begin{aligned}\frac{n_{+.}}{q_{10}} - \frac{n_{++}}{1-q_{10}-q_{01}} &= 0,\\ \frac{n_{.+}}{q_{01}} - \frac{n_{++}}{1-q_{10}-q_{01}} &= 0,\\ \frac{n_{1.}}{p_{1.}} - \frac{n_{0.}}{1-p_{1.}} &= 0,\\ \frac{n_{.1}}{p_{.1}} - \frac{n_{.0}}{1-p_{.1}} &= 0,\\ \frac{n_{10}}{p_{10}} - \frac{n_{00}}{1-p_{10}-p_{01}-p_{11}} &= 0,\\ \frac{n_{01}}{p_{01}} - \frac{n_{00}}{1-p_{10}-p_{01}-p_{11}} &= 0,\\ \frac{n_{11}}{p_{11}} - \frac{n_{00}}{1-p_{10}-p_{01}-p_{11}} &= 0.\end{aligned}$$

Iš pirmos lygties

$$n_{+.}(1-q_{10}-q_{01}) = n_{++}q_{10},$$

$$q_{01} = 1 - q_{10} - \frac{n_{++}}{n_{+.}}q_{10}.$$

Įstatę į antrą lygybę gauname

$$\frac{n_{.+}}{1-q_{10}-\frac{n_{++}}{n_{+.}}q_{10}} - \frac{n_{++}}{\frac{n_{++}}{n_{+.}}q_{10}} = 0,$$

$$\frac{n_{.+}}{1-q_{10}-\frac{n_{++}}{n_{+.}}q_{10}} - \frac{n_{+.}}{q_{10}} = 0,$$

$$q_{10} = \frac{n_{+.}}{n_{++} + n_{.+} + n_{+.}}.$$

Taigi

$$\hat{q}_{10} = \frac{n_{+.}}{n},$$

$$\hat{q}_{01} = 1 - \frac{n_{+.}}{n} - \frac{n_{++}}{n_{+.}}\frac{n_{+.}}{n} = \frac{n_{.+}}{n}.$$

Iš 3 ir 4 lygčių iškart gaunami parametrų $p_{1.}$ ir $p_{.1}$ įverčiai:

$$\hat{p}_{1.} = \frac{n_{1.}}{n_{+.}},$$

$$\hat{p}_{.1} = \frac{n_{.1}}{n_{.+}}.$$

Likusius įverčius gausime iš paskutinių trijų lygčių.

Pažymime

$$c = \frac{n_{00}}{1 - p_{10} - p_{01} - p_{11}},$$

tada

$$c = \frac{n_{10}}{p_{10}} = \frac{n_{01}}{p_{01}} = \frac{n_{11}}{p_{11}}.$$

Iš čia

$$p_{10} = \frac{n_{10}}{c}, \quad p_{01} = \frac{n_{01}}{c}, \quad p_{11} = \frac{n_{11}}{c}, \quad 1 - p_{10} - p_{01} - p_{11} = \frac{n_{00}}{c}.$$

Sudėję visas lygybes, gauname

$$1 = \frac{n_{00} + n_{10} + n_{01} + n_{11}}{c} = \frac{n_{++}}{c},$$

t.y. $n = c$. Taigi

$$\hat{p}_{10} = \frac{n_{10}}{n_{++}},$$

$$\hat{p}_{01} = \frac{n_{01}}{n_{++}},$$

$$\hat{p}_{11} = \frac{n_{11}}{n_{++}}.$$

2.2.3 Vienos hipotezės tikrinimas

Iš (Y, X) skirstinio aprašymo matyti, kad $p_{10} + p_{11}$ yra tikimybė nupirkti 1 prekę, kai buvo pasiūlytos abi prekės, o $p_{1.}$ – ta pati tikimybė, kai antroji prekė nebuvo siūlyta. Logiška, kad abi tikimybės turėtų sutapti. Sukonstruosime kriterijų hipotezei $H : p_{10} + p_{11} = p_{1.}$ tikrinti. Tam rasime dydžio $(\hat{p}_{10} + \hat{p}_{11} - \hat{p}_{1.})$ asimptotinį skirstinį.

Aišku, kad

$$\begin{aligned} T &= \frac{\overline{1_{1110}} + \overline{1_{1111}}}{\overline{1_{1100}} + \overline{1_{1101}} + \overline{1_{1110}} + \overline{1_{1111}}} - \frac{\overline{1_{1010}}}{\overline{1_{1000}} + \overline{1_{1010}}} \\ &= g(\overline{1_{1110} + 1_{1111}}, \overline{1_{1100} + 1_{1101}}, \overline{1_{1010}}, \overline{1_{1000}}), \end{aligned}$$

čia

$$g(a, b, c, d) = \frac{a}{a+b} - \frac{c}{c+d},$$

$1_{u_1u_2u_3u_4} = 1$, kai $(X, Y) = (u_1, u_2, u_3, u_4)$ ir 0 kitais atvejais, o brūkšnys žymi empirinį vidurkį.

Iš CRT:

$$\sqrt{n}\left[\begin{pmatrix} \overline{1_{1110} + 1_{1111}} \\ \overline{1_{1100} + 1_{1101}} \\ \overline{1_{1010}} \\ \overline{1_{1000}} \end{pmatrix} - \begin{pmatrix} q_{11}p_{1+} \\ q_{11}p_{0+} \\ q_{10}p_{1.} \\ q_{10}p_{0.} \end{pmatrix}\right] \to N(0, \Sigma);$$

čia

$$\Sigma = \begin{pmatrix} q_{11}p_{1+} - q_{11}^2p_{1+}^2 & -q_{11}^2p_{1+}p_{0+} & -q_{11}p_{1+}q_{10}p_{1.} & -q_{11}p_{1+}q_{10}p_{0.} \\ -q_{11}^2p_{1+}p_{0+} & q_{11}p_{0+} - q_{11}^2p_{0+}^2 & -q_{11}p_{0+}q_{10}p_{1.} & -q_{11}p_{0+}q_{10}p_{0.} \\ -q_{11}p_{1+}q_{10}p_{1.} & -q_{11}p_{0+}q_{10}p_{1.} & q_{10}p_{1.} - q_{10}^2p_{1.}^2 & -q_{10}p_{1.}q_{10}p_{0.} \\ -q_{11}p_{1+}q_{10}p_{0.} & -q_{11}p_{0+}q_{10}p_{0.} & -q_{10}p_{1.}q_{10}p_{0.} & q_{10}p_{0.} - q_{10}^2p_{0.}^2 \end{pmatrix}$$

Panaudoję delta metodą gauname $\sqrt{n}(T - g_0) \to g_0' N(0, \Sigma)$, čia

$$\begin{aligned} g_0 &= g(q_{11}p_{1+}, q_{11}p_{0+}, q_{10}p_{1.}, q_{10}p_{0.}) \\ g_0' &= g'(q_{11}p_{1+}, q_{11}p_{0+}, q_{10}p_{1.}, q_{10}p_{0.}) \end{aligned}$$

ir

$$g' = \left(\tfrac{\partial g}{\partial a}\, \tfrac{\partial g}{\partial b}\, \tfrac{\partial g}{\partial c}\, \tfrac{\partial g}{\partial d}\right)$$

Mūsų atveju

$$g_0 = \frac{q_{11}p_{1+}}{q_{11}p_{1+} + q_{11}p_{0+}} - \frac{q_{10}p_{1.}}{q_{10}p_{0.} + q_{10}p_{1.}} = p_{1+} - p_{1.} = 0,$$

(jei H teisinga),

$$g'(a,b,c,d) = \left(\begin{array}{cccc} \frac{b}{(a+b)^2} & \frac{-a}{(a+b)^2} & \frac{-d}{(c+d)^2} & \frac{c}{(c+d)^2} \end{array} \right)$$

$$g_0' = \left(\begin{array}{cccc} \frac{p_{0+}}{q_{11}} & \frac{-p_{1+}}{q_{11}} & \frac{-p_{0.}}{q_{10}} & \frac{p_{1.}}{q_{10}} \end{array} \right)$$

Taigi

$$\sqrt{n}(T-0) \to N(0,\sigma^2)$$

su

$$\begin{aligned} \sigma^2 &= \left(\begin{array}{cccc} \frac{p_{0+}}{q_{11}} & \frac{-p_{1+}}{q_{11}} & \frac{-p_{0.}}{q_{10}} & \frac{p_{1.}}{q_{10}} \end{array} \right) \Sigma \begin{pmatrix} \frac{p_{0+}}{q_{11}} \\ \frac{-p_{1+}}{q_{11}} \\ \frac{-p_{0.}}{q_{10}} \\ \frac{p_{1.}}{q_{10}} \end{pmatrix} \\ &= \left(\begin{array}{cccc} \frac{p_{0+}}{q_{11}} & \frac{-p_{1+}}{q_{11}} & \frac{-p_{0.}}{q_{10}} & \frac{p_{1.}}{q_{10}} \end{array} \right) \begin{pmatrix} p_{1+}p_{0+} \\ -p_{0+}p_{1+} \\ -p_{0.}p_{1.} \\ p_{0.}p_{1.} \end{pmatrix} \\ &= \frac{p_{0+}^2 p_{1+}}{q_{11}} + \frac{p_{0+}p_{1+}^2}{q_{11}} + \frac{p_{0.}^2 p_{1.}}{q_{10}} + \frac{p_{0.}p_{1.}^2}{q_{10}} \\ &= \frac{p_{0+}p_{1+}}{q_{11}} + \frac{p_{0.}p_{1.}}{q_{10}}. \end{aligned}$$

Jei pažymėsime

$$\widehat{\sigma} = \sqrt{\frac{\hat{p_{0+}}\hat{p_{1+}}}{\hat{q_{11}}} + \frac{\hat{p_{0.}}\hat{p_{1.}}}{\hat{q_{10}}}} = \sqrt{\frac{\frac{n_{0+}}{n_{++}}\frac{n_{1+}}{n_{++}}}{\frac{n_{++}}{n}} + \frac{\frac{n_{0.}}{n_{+.}}\frac{n_{1.}}{n_{+.}}}{\frac{n_{+.}}{n}}},$$

tai $\hat{\sigma} \to \sigma$ ir $\frac{\sqrt{n}T}{\hat{\sigma}} \to N(0,1)$ (jei hipotezė teisinga).

Todėl hipotezė atmetama, kai $|\tilde{T}| > z_{\alpha/2}$, čia

$$|\tilde{T}| = \frac{|T|}{|\sqrt{\frac{n_{0+}n_{1+}}{n_{++}^3} + \frac{n_{0.}n_{1.}}{n_{+.}^3}}|},$$

$z_{\alpha/2}$ žymi normaliojo skirstinio kritinę reikšmę, o α yra reikšmingumo lygmuo.

Patikrinsime hipotezę su prekėmis 1 ir 2. Mūsų atveju

$$n_{10} = 16, \quad n_{11} = 225, \quad n_{++} = 298, \quad n_{1.} = 445, \quad n_{+.} = 557,$$

$$n_{0+} = 57, \quad n_{1+} = 241, \quad n_{0.} = 112,$$

todėl $|\tilde{T}|$ yra lygi 0,4, o $z_{0,025} = 1,96$ (kai $\alpha = 0,05$). Gauname, kad hipotezė yra neatmetama su reikšmingumo lygmeniu $\alpha = 0,05$, nes $|\tilde{T}| < z_{0,025}$. Taigi galime teigti, kad tikimybė, kad klientas pirks pirmą prekę nepriklauso nuo to, ar antra prekė buvo pasiūlyta klientui.

2.3 Pirmas modelis

2.3.1 Klasikinė situacija

Paprastai imties tūris n būna nepakankamai didelis, kad būtų galima įvertinti visus $2^N - 1$ parametrus. Todėl kuriami atsitiktinio pirkinių krepšelio modeliai, aprašomi mažesniu skaičiumi parametrų. Paprasčiausias toks modelis remiasi prielaida, kad pirkiniai į krepšelį dedami nepriklausomai vienas nuo kito, t.y. kad $B(1), \ldots, B(N)$ dydžiai nepriklausomi. Tokiu atveju

$$p(b) = \prod_{i \in b} q_i \prod_{i \notin b} (1 - q_i);$$

čia $q_i = \mathrm{P}\{B(i) = 1\}$ yra modelio parametrai. Taigi vietoje $2^N - 1$ parametro šiame modelyje yra tik N parametrų (jei $N = 4$, turime tik 4 parametrus vietoje 15).

Patikrinkime, ar šis modelis tinka mūsų duomenims, panaudoję tikėtinumo santykio kriterijų. Kriterijaus statistika $Q = 2[\ell(\hat{p}) - \ell(\tilde{p})]$, čia $\hat{p}$ – parametro didžiausio tikėtinumo įvertinys "pilname" modelyje, $\tilde{p}$ – didžiausio tikėtinumo įvertinys "susiaurintame" modelyje (t.y. laikant, kad hipotezė teisinga), o $\ell(p)$ - tikėtinumo funkcija. Jei hipotezė teisinga, tai $Q \to \chi^2(d)$, su $d = 2^N - 1 - N$. Taigi hipotezė atmetama, jei

$$Q > \chi^2_\alpha(2^N - 1 - N).$$

Aišku, kad $\tilde{p}(b) = \prod_{i \in b} \hat{q}_i \prod_{i \notin b} (1 - \hat{q}_i)$, čia $\hat{q}_i$ yra q_i reikšmės, su kuriomis maksimali funkcija

$$\tilde{L} = \prod_{k=1}^{n} \Big(\prod_{i \in b_k} q_i \prod_{i \notin b_k} (1 - q_i) \Big).$$

Aišku, kad

$$\tilde{L} = \prod_{i=1}^{N} q_i^{\tilde{n}_i} (1 - q_i)^{n - \tilde{n}_i};$$

čia $\tilde{n}_i$ – skaičius krepšelių b_k, kuriuose yra i-oji prekė.

Tikėtinumo funkcijos logaritmas

$$\tilde{\ell} = \sum_{i=1}^{N} (\tilde{n}_i \ln q_i + (n - \tilde{n}_i) \ln(1 - q_i)),$$

Suskaičiuojame išvestines ir prilyginame 0; tada

$$\frac{\tilde{n}_i}{q_i} - \frac{n - \tilde{n}_i}{1 - q_i} = 0,$$
$$\tilde{n}_i(1 - q_i) = (n - \tilde{n}_i)q_i,$$
$$\tilde{n}_i = nq_i,$$

ir

$$\hat{q}_i = \frac{\tilde{n}_i}{n}.$$

Patikrinsime, ar modelis tinka mūsų duomenims, kai buvo siūlytos visos keturios prekės. Suskaičiuojame q_i įverčių reikšmes "pilno" ir "susiaurinto" modelių atvejais. Mūsų atveju:

$$n_\emptyset = 6, \quad n_4 = 2, \quad n_{24} = 1, \quad n_{34} = 1, \quad n_{123} = 3,$$
$$n_{124} = 7, \quad n_{134} = 2, \quad n_{234} = 5, \quad n_{1234} = 51, \quad n = 78.$$

"Susiaurinto" modelio tikėtinumo funkcijos $l(\tilde{p})$ logaritmo reikšmė yra $-137,39$. "Pilno" modelio tikėtinumo funkcijos logaritmo reikšmė (čia $\hat{p}(b) = \frac{n(b)}{n}$) yra $l(\hat{p}) = -100,82$. Gauname, kad hipotezė yra atmetama, nes kriterijaus statistika $Q = 73,14$ yra daugiau už $\chi^2_\alpha(11)$ reikšmę 19,68. Taigi šis modelis netinka duomenims, nors reikia atsižvelgti į tai, kad imtį sudaro tik 78 krepšeliai.

2.3.2 Netipinė situacija

Patikrinkime, ar pirmas modelis tinka mūsų duomenims, kai atsižvelgiame į tai, kad pirkėjui buvo pasiūlytos ne visos prekės (imsime tik dvi prekes: 1 ir 2). Vėl naudosime tikėtinumo santykio kriterijų. Tarkime, turime imtį iš skirstinio aprašyto 2.2.2 skyrelyje (kai tiriame dvi prekes). Jo parametrų aibė Θ susideda iš visų

$$\theta = (q_{10}, q_{01}, 1 - q_{10} - q_{01}, p_{1\cdot}, 1 - p_{1\cdot}, p_{\cdot 1}, 1 - p_{\cdot 1}, p_{01}, p_{10}, p_{11}, 1 - p_{01} - p_{10} - p_{11})$$

tenkinančių sąlygas

$$q_{10}, q_{01} > 0, \quad q_{10} + q_{01} < 1,$$
$$0 < p_{1\cdot} < 1, \quad 0 < p_{\cdot 1} < 1,$$

$$p_{10}, p_{01}, p_{11} > 0,$$
$$p_{10} + p_{01} + p_{11} < 1,$$

Taikomą modelį atitinka tam tikras parametrų aibės poaibis $M \subset \Theta$.

Kadangi laikome, kad $B(1), \ldots, B(N)$ dydžiai nepriklausomi, $\theta \in M$, kai su tam tikrais $q_1, q_2 \in (0,1)$

$$p_{1.} = q_1,$$
$$p_{.1} = q_2,$$
$$p_{0.} = 1 - q_1,$$
$$p_{.0} = 1 - q_2,$$
$$p_{00} = (1 - q_1)(1 - q_2),$$
$$p_{10} = q_1(1 - q_2),$$
$$p_{01} = (1 - q_1)q_2,$$
$$p_{11} = q_1 q_2.$$

Didžiausio tikėtinumo įverčiai "susiaurintame" modelyje skaičiuojami pagal tokias pat formules, tik su $\hat{q}_i$ vietoje q_i, čia $\hat{q}_1$, $\hat{q}_2$ yra reikšmės su kuriomis didžiausią reikšmę įgyja funkcija

$$\begin{aligned}\tilde{\ell} = n_{+.} \ln q_{10} + n_{.+} \ln q_{01} + n_{++} \ln(1 - q_{10} - q_{01}) + \\ + (n_{1.} + n_{10} + n_{11}) \ln q_1 + (n_{0.} + n_{00} + n_{01}) \ln(1 - q_1) + (n_{.1} + \\ + n_{01} + n_{11}) \ln q_2 + (n_{.0} + n_{00} + n_{10}) \ln(1 - q_2).\end{aligned}$$

Aišku, kad

$$\frac{\partial \tilde{\ell}}{\partial q_1} = \frac{n_{1.} + n_{10} + n_{11}}{q_1} - \frac{n_{01} + n_{0.} + n_{00}}{1 - q_1},$$

$$\frac{\partial \tilde{\ell}}{\partial q_2} = \frac{n_{.1} + n_{01} + n_{11}}{q_2} - \frac{n_{10} + n_{.0} + n_{00}}{1 - q_2}.$$

Prilyginę išvestines 0, gauname parametrų q_1 ir q_2 įverčius:

$$\hat{q}_1 = \frac{n_{1.} + n_{10} + n_{11}}{n_{1.} + n_{10} + n_{11} + n_{01} + n_{0.} + n_{00}},$$
$$\hat{q}_2 = \frac{n_{.1} + n_{01} + n_{11}}{n_{.1} + n_{01} + n_{11} + n_{10} + n_{.0} + n_{00}}.$$

Parametrų q_{10}, q_{01} įverčiai tokie patys, kaip skyrelyje 2.2.2.

Patikrinsime, ar modelis tinka mūsų duomenims, kai buvo siūlytos dvi prekės. Suskaičiuojame q_i įverčių reikšmes "pilno" ir "susiaurinto" modelių atvejai. Mūsų atveju:

$$n_{0\cdot} = 112, \quad n_{\cdot 0} = 489, \quad n_{\cdot 1} = 2665, \quad n_{1\cdot} = 445,$$
$$n_{10} = 16, \quad n_{01} = 36, \quad n_{11} = 225, \quad n_{00} = 21.$$

"Susiaurinto" modelio tikėtinumo funkcijos $l(\tilde{p})$ logaritmo reikšmė yra $-4530,31$. "Pilno" modelio tikėtinumo funkcijos logaritmo reikšmė $l(\hat{p}) = -4512,41$. Gauname, kad hipotezė yra atmetama, nes kriterijaus statistika $Q = 35,8$ yra daugiau už $\chi^2_\alpha(3)$ reikšmę 7. Taigi, šis modelis netinka duomenims.

2.4 Antras modelis

Dabar panagrinėsiu modelį, aprašytą [3] straipsnyje ir pritaikysiu mūsų turimiems duomenims. Jame yra N parametrų β_i, $1 \le i \le N$, ir dar $\binom{N}{2}$ parametrų θ_{ij}, $1 \le i < j \le N$, o krepšelio tikimybė

$$p(b_k) = \frac{\mathrm{e}^{\mu(b_k)}}{\sum_{b'_k} \mathrm{e}^{\mu(b'_k)}};$$

čia

$$\mu(b_k) = \sum_{i \in b_k} \beta_i + \sum_{\substack{i,j \in b_k \\ i<j}} \theta_{ij}.$$

Norint rasti parametrų didžiausio tikėtinumo įverčius reikia maksimizuoti tikėtinumo funkciją

$$\prod_{k=1}^{n} p(b_k) = \prod_{k=1}^{n} \frac{\mathrm{e}^{\mu(b_k)}}{\sum_{b'_k} \mathrm{e}^{\mu(b'_k)}},$$

Tikėtinumo funkciją maksimizuoti galime ir su SAS procedūra PHREG (kaip panaudoti procedūrą apibendrintai logistinei regresijai aprašyta [4] straipsnyje). Naudosime tik tuos duomenis, kai klientams buvo pasiūlytos visos 4 prekės. Gauti rezultatai

```
                Analysis of Maximum Likelihood Estimates

             Parameter Standard Chi-              Hazard Variable
Variable DF Estimate  Error    Square   Pr>ChiSq Ratio  Label
```

a1	1	-3.61547	1.98011	3.3339	0.0679	0.027	a1
a2	1	-0.69287	1.22464	0.3201	0.5715	0.500	a2
a3	1	-0.69301	1.22469	0.3202	0.5715	0.500	a3
a4	1	-1.09861	0.81650	1.8104	0.1785	0.333	a4
a1a2	1	1.62928	1.31127	1.5439	0.2140	5.100	a1a2
a1a3	1	0.37675	1.16713	0.1042	0.7468	1.458	a1a3
a1a4	1	3.93183	1.00976	15.1618	<.0001	51.000	a1a4
a2a3	1	2.30217	1.64305	1.9632	0.1612	9.996	a2a3
a2a4	0	0	.	.	.	.	a2a4
a3a4	0	0	.	.	.	.	a3a4

Kintamieji a1-a4 atitinka prekes $1-4$, o kintamieji a1a2-a3a4 atitinka dviejų prekių poveikį viena kitai. Šiuo atveju gauname, kad statistiškai reikšmingas yra tik prekių 1 ir 4 ryšys, kadangi gauname teigiamą parametro įvertinį, šios prekės papildo vieną kitą.

2.5 Trečias modelis

Šis modelis buvo sugalvotas taip, kad tiktų prekių pardavinėjimui telefonu, t.y. kai vieno skambučio metu nuperkama viena prekė ir pirkėjui gali būti nepasiūlytos visos prekės.

Laikysime, kad kliento poreikis pirkti prekę yra proporcingas jo norui pirkti prekę ir atvirščiai proporcingas pinigų sumai išleistai per paskutinį pirkinį. Taigi tariame, kad k-asis klientas pirks prekę, jei

$$\frac{N_k}{y_k} \geq 1, \tag{1}$$

čia N_k žymime kliento norą įsigyti prekę, o y_k yra jau siūlytos prekės kaina.

Tarkime, kliento noras pirkti prekę nukrenta iki 0 po kiekvieno eilinio pirkimo ir po to auga tiesiškai iki kito pirkinio, t.y. $N_k = A_k t_k$, čia t_k – laikas, praėjęs nuo paskutinio pirkimo, o A_k – tam tikras atsitiktinis dydis. Tegu Z_k yra atsitiktinis dydis, lygus 1, jei klientas nupirko prekę, ir lygus 0, jei nenupirko. Tada $Z_k = 1$, kai

$$\frac{A_k t_k}{y_k} \geq 1,$$
$$\ln A_k + \ln t_k - \ln y_k \geq 0,$$

Laikykim, kad $\ln A_k$ yra nepriklausomi atsitiktiniai dydžiai, pasiskirstę normaliai su nežinomais vidurkiais μ_k ir tam tikra pastovia dispersija σ^2, o vidurkis

$$\mu_k = a + b \ln d_k,$$

čia d_k – nupirktų prekių skaičius. Tada

$$\ln A_k = \mu_k + \sigma^2 \epsilon_k = a + b \ln d_k + \sigma^2 \epsilon_k$$

(ϵ_k žymi nepriklausomus atsitiktinius dydžius su standartiniu normaliu skirstiniu) ir

$$\begin{aligned} \mathrm{P}\{Z_k = 1\} &= \mathrm{P}\Big\{a + b \ln d_k + \sigma^2 \epsilon_k + \ln t_k - \ln y_k \geq 0\Big\} \\ &= \mathrm{P}\Big\{\epsilon_k > \frac{-a - b \ln d_k - \ln t_k + \ln y_k}{\sigma}\Big\} \\ &= 1 - \Phi\Big(\frac{-a - b \ln d_k - \ln t_k + \ln y_k}{\sigma}\Big). \end{aligned}$$

Norint įvertinti nežinomus parametrus a, b ir σ reikia maksimizuoti tikėtinumo funkciją

$$\tilde{L} = \prod_{k=1}^{n} (1 - \Phi(\frac{-a - b \ln d_k - \ln t_k + \ln y_k}{\sqrt{\sigma^2}}))$$

Tikėtinumo funkcijos maksimizavimui naudosime SAS procedūrą LOGISTIC. Tirsime keturis produktus 1, 2, 3 ir 4 ir tikimybę, kad klientas nupirko prekę 4. Kintamasis `tk` žymės intervalą tarp laiko, kada buvo siūlyta prekė 4 ir laiko, kada buvo nupirkta prieš tai siūlyta prekė, šios prekės kainą žymės kintamasis `yk`. O kintamasis `dk` – tai kliento nupirktų prekių skaičius. Gauname parametrų įverčius:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9304	2.9521	0.4276	0.5132
lnyk	1	-0.0668	0.5688	0.0138	0.9065
lntk	1	0.1408	0.1768	0.6342	0.4258
lndk	1	3.7807	0.3768	100.6892	<.0001

Pats modelis neblogai tinka duomenims (Max-rescaled R-Square reikšmė lygi 0.6294).

Paskaičiuojame parametrų įverčius iš modelio pašalinę kainos logaritmą.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2571	0.9503	5.6415	0.0175
lntk	1	0.1483	0.1637	0.8204	0.3651
lndk	1	3.7807	0.3768	100.6719	<.0001

Modelis taip pat tinka duomenims (Max-rescaled R-Square reikšmė lygi 0.6294). Gauname, kad kliento poreikiui pirkti prekę daugiausiai įtakos turi tai, kiek jis prekių yra pirkęs, t.y. kuo daugiau klientas yra nupirkęs, tuo didesnė tikimybė, kad jis pirks ir toliau.

Įvertinsime parametrus a, b ir σ. Gauname, kad

$$\frac{a}{\sigma} = -2.2571,$$
$$\frac{b}{\sigma} = 3.7807,$$
$$\frac{1}{\sigma} = 0.1483.$$

Iš čia gauname, kad kliento noras pirkti prekę yra pasiskirstęs normaliai su parametrais $(-15.22 + 25.49 \ln d_k,\ 6.74)$.

Išvados

Gauti rezultatai tiriant laiko tarp skambučių ir mokėjimų intervalus parodė, kad klientai yra linkę pirkti prekes dažniau nei jiems yra skambinama. Dažniausiai klientai yra nupirkę tik vieną prekę iš tiriamų keturių.

Kadangi nevisiems klientams buvo siūlytos visos prekės, aprašėme tokių netipinių krepšelių skirstinį. Naudodami didžiausio tikėtinumo metodą įvertinome parametrus klasikinių ir netipinių krepšelių atvejais. Iškėlėme hipotezę, kad $p_{10} + p_{11} = p_{1.}$, ir išvedėme kriterijų jai tikrinti, mūsų naudojamiems duomenims hipotezė buvo neatmesta.

Išnagrinėjome modelį su prielaida, kad prekės į krepšelius yra dedamos atsitiktinai. Panaudoję tikėtinumo santykio kriterijų patikrinome, ar mūsų duomenims tinka šis modelis klasiniu ir netipinių krepšelių situacijose. Abiem atvejais gavome, kad modelis duomenims netinka.

Antrame modelyje patikrinome, kaip kliento turimas krepšelis priklauso nuo pasirinktų prekių ir jų tarpusavio ryšių. Gavome, kad reikšmingas yra tik knygų 1 ir 4 ryšys.

Trečiame modelyje gavome, kad prekės 1 pirkimas labiausiai priklauso nuo to, kiek klientas yra pirkęs prekių ir nuo laiko, kada buvo siūlyta paskutinė prekė. Taip pat įvertinome kliento noro pirkti skirstinio parametrus.

Aprašytus modelius galima pritaikyti įvairių tipų krepšelių duomenims analizuoti. Darbą galima toliau plėsti tiriant kitus modelius, iškeliant naujas hipotezes dėl prekių pasirinkimo.

Santrauka

Kai prekės parduodamos telefonu, įprasti pirkinių krepšelio modeliai netinka, nes ne visos prekės būna pasiūlytos visiems klientams. Šiame magistro darbe keli žinomi modeliai adaptuoti tokiai situacijai. Be to, sukurtas naujas modelis, kuriame atsižvelgiama į kliento pirkimų apimtis ir laiką, praėjusį nuo paskutinio pirkimo. Visi modeliai pritaikyti realiems pardavimų duomenims.

Summary

When items are sold via phone, common market basket models can not be applied, because not every item has been proposed to every customer. In this master thesis some known models are adapted to this particular case. Moreover, the new model takes the amount of customer purchases' and the time since his last purchase into consideration. All models are applied to real trade data.

Priedai

SAS procedūros naudotos pirmame skyriuje ([2])

```
goptions
hsize=14in vsize=12in
border htext=12pt;
PROC GCHART DATA=D2;
HBAR total / TYPE =PERCENT DISCRETE;
RUN;Quit;
PROC GCHART DATA=D4;
HBAR zi /group=prod DISCRETE SPACE=0 WIDTH=1.5;
RUN; QUIT;
PROC UNIVARIATE DATA=D;
VAR ti;
HISTOGRAM ti /MIDPOINTS=30 TO 630 BY 30;
RUN;
PROC univariate DATA=d CIBASIC CIPCTLNORMAL(ALPHA=0.05);
VAR ti;
RUN
```

SAS programėlė skyrelio 2.2.3 hipotezei tikrinti:

```
%LET P1 = 'a1';
%LET P2 = 'a2';
%LET alfa = 0.05;
Proc sql;
create table sk2 as

SELECT * FROM
(select count(*) as n0p from d a, d b
WHERE a.prod2 = &P1 AND b.prod2 = &P2 AND a.code = b.code and a.zi=0),

(select count(*) as n1p from d a, d b
WHERE a.prod2 = &P1 AND b.prod2 = &P2 AND a.code = b.code and a.zi=1),

(select count(*) as npp from d a, d b
WHERE a.prod2 = &P1 AND b.prod2 = &P2 AND a.code = b.code),

(select count(distinct code) as n0 FROM d
WHERE prod2 = &P1 AND zi = 0),
```

```
(select count(distinct code) as n1 FROM d
WHERE prod2 = &P1 AND zi = 1),

(select count(distinct a.code) as np FROM d a, d b
WHERE a.prod2 NOT IN ('a2', 'a3', 'a4') AND b.prod2 NOT IN ('a2', 'a3', 'a4') AN

(select count(*) as n10 from d a, d b
WHERE a.prod2 = &P1 AND b.prod2 = &P2 AND a.code = b.code and a.zi=1 and b.zi =

(select count(*) as n11 from d a, d b
WHERE a.prod2 = &P1 AND b.prod2 = &P2 AND a.code = b.code and a.zi=1 and b.zi =
run;
data t;
set sk2;
np=np-npp;
n0=n0-n0p;
n1=n1-n1p;
sigma=sqrt((n0p*n1p)/(npp*npp*npp)+(n0*n0)/(np*np*np));
t=((n10/npp)+(n11/npp)-(n1/np))/sigma;
krit_r=PROBIT(1-&alfa/2);
run;
```

SAS procedūra PHREG antram modeliui:

```
data choice2;
set d;
array inters[4]
a1 /*Pirmoji preke */
a2 /*Antroji preke */
a3 /*Trecioji preke */
a4; /*Ketvirtoji preke */
array cross[6]
a1a2 /*rysys tarp pirmos ir antros prekes */
a1a3 /*rysys tarp pirmos ir trecios prekes */
a1a4 /*rysys tarp pirmos ir ketvirtos prekes */
a2a3 /*rysys tarp antros ir trecios prekes */
a2a4 /*rysys tarp antros ir ketvirtos prekes */
a3a4; /*rysys tarp trecios ir ketvirtos prekes */
k=1;
do i = 1 to 4;
```

```
do j = 1 to 4;
if j>i then do;
cross[k]=inters[i]*inters[j]*inters[i];
k=k+1;
end;
end;
end; drop F8-F13; output; run;

proc phreg data=choice2;
model pas*pas(2) = a1 a2 a3 a4 a1a2 a1a3 a1a4 a2a3 a2a4 a3a4/ ties=breslow;
/* naudojame Breslow tiketinumo funkcija */
strata code;
run;
```

SAS procedūra LOGISTIC trečiam modeliui:

```
data d2;
set d;
lnyk=log(yk);
lntk=log(tk);
lndk=log(dk);
run;

proc logistic data=d2;
model zk(event='1')= lnyk lntk lndk / link=probit rsq;
run;
```

Literatūra

[1] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD Conference*, pages 265–276, 1997.

[2] Rūta Levulienė. *Statistikos taikymai naudojant SAS@.* VUL, 2009.

[3] Gary J. Russell and Ann Petersen. Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76(3):367–392, 2000.

[4] Ying So and Warren F. Kuhfeld. Multinomial logit models. In *SUGI 20*, pages 665–680, 2010.