# A DISZKRÉT KIVÁLASZTÁSI MODELL BECSLÉSE COX-REGRESSZIÓVAL[1]

HAJDU OTTÓ
*BME GTK*

## 1 Bevezetés

A tanulmány az MDC (multinomial discrete choice) kiválasztási modell paramétereinek a standard pontbecslési eljárásait tekinti át egyfelől az eredmények értelmezését illetően, másfelől a modell gyakorlati alkalmazását segítendő azon esetre, mikor közvetlenül működő standard MDC programcsomag nem áll rendelkezésre.[2] A probléma klasszikus megoldása az irodalomban a túlélési modellek egyik alapmódszerének, az ún. Cox-regressziónak a felhasználásával történik, amelynek rugalmas, kontrollálható alkalmazása akkor is előnyt nyújthat, ha egyébként standard MDC program is rendelkezésre áll.[3]

Az MDC modellben egy $I$ individuum (a döntést hozó személy) $g = 1, 2, \ldots, m_I$ számú lehetséges *alternatíva* halmazából egyet biztosan, és csak egyet kiválaszt. A választási alternatívák száma individuumonként nem szükségszerűen azonos. Nyilvánvaló, hogy a kiválasztás eredményét mind a döntéshozó individuális tulajdonságai, mind az alternatíva sajátosságai befolyásolják. A döntés kimenetét (eredményét) magyarázó változók tehát egyfelől lehetnek *individuális* jellegűek, másfelől *alternatíva-specifikusak*. Az előbbieket $x$, az utóbbiakat $Z$ jelöli a tanulmányban. Bár a modell paramétereinek a becslése mindkét típus esetén a maximum likelihood (ML) elven alapul, jellegét tekintve különbözik $x$, vagy $Z$ típusú magyarázó változók használata esetén. Előbb e két almodell becslésének a különbözősége kerül tárgyalásra. Ezt követően, mivel a reális döntési helyzet mind $x$, mind $Z$ egyidejű figyelembe vételét („vegyes" modell alkalmazását) igényli, a becslési eljárást az általános, mindkét típust magában foglaló esetre is bemutatjuk.

Az alternatíva-specifikus magyarázó változók kezelése, és így a „vegyes" modell becslése standard módon a statisztikai szoftverek többségében közvetlenül nem érhető el. Ennek okán a tanulmány —egy illusztratív példa számszerűsítésén át— útmutatást ad arra vonatkozóan, hogy a „vegyes" modell adatállományát milyen struktúrában kell rögzíteni (a modellt hogyan kell paraméterezni) annak érdekében, hogy paramétereinek a becslése a standard *Cox-regresszióval* megoldhatóvá váljon.

[1]Beérkezett: 2005. június 21. E-mail: ohajdu@finance.bme.hu

[2]Például a Systat 11.0, vagy a SAS program szolgáltat közvetlenül hívható MDC modult, viszont például az SPSS 13.0 programban az eljárás csak közvetetten oldható meg.

[3]Lásd például Kuhfeld (2003), vagy a SAS program MDC és QLIM eljárásait.

## 2 A polichotom logit modell esete

A polichotom (multinomiális) logit modell (PL) akkor alkalmazandó, ha a döntés diszkrét kimenetét (a kiválasztás eredményét) magyarázó változó *individuális* jellegű, és az alternatívák köre mindenkire azonos: $g = 1, 2, \ldots, m$. Jelölje $x_I$ a döntést hozó $I$ individuum valamely jellemzőjét: példánkban az életkorát. Ekkor annak a valószínűsége, hogy a választható alternatívák közül éppen a $C \in \{1, 2, \ldots, m\}$ alternatívát választja ki az $I$ személy:

$$P_{CI} = \frac{P_{CI}}{\sum_{g=1}^{m} P_{gI}} = \frac{P_{CI}/P_{MI}}{\sum_{g=1}^{m} P_{gI}/P_{mI}} = \frac{\text{odds}_I(C:m)}{\sum_{g=1}^{m} \text{odds}_I(g:m)} = \frac{e^{\alpha_C + \beta_C x_I}}{\sum_{g=1}^{m} e^{\alpha_g + \beta_g x_I}} ,$$

ahol az $\text{odds}_I(C:m)$ valószínűségarány annak az esélye, hogy az $I$ személy a $C$ alternatívát preferálja az $m$ referencia alternatívával szemben ($\text{odds}_I(m:m) = 1$), és az $\ln(\text{odds}) = \textit{logit}$ mennyiség a magyarázó változó lineáris függvénye az $\alpha_m = \beta_m = 0$ megkötés mellett.

Ha a döntéshozó előtt csak két alternatíva áll, akkor $m = 2$ mellett az ún. *dichotom*, vagy bináris logit modellt kapjuk.

A modell szerint mind az $\alpha_g$ tengelymetszet, mind a $\beta_g$ parciális regressziós paraméter *alternatíva-specifikus*, miközben adott individuum $x_I$ jellemzője ugyanaz bármely alternatíva esetén. Látható, hogy $m$ számú alternatíva mellett a paraméterek $m-1$ számú körét definiáljuk. Ha az $x$ változó értéke egységnyit emelkedik, akkor a $C:m$ viszonylatú *odds* az $e^{\beta_C}$ *odds-ratio* (OR) faktorral szorzódik.

A regressziós paramétereket kézenfekvő a maximum likelihood (ML) módszerrel becsülni. Jelölje a $C_1, C_2, \ldots, C_n$ minta $I = 1, 2, \ldots, n$ individuum megfigyelt, független döntéseit: $C_I \in \{1, 2, \ldots, m\}$. A minta maximálandó likelihood függvénye:

$$L = \prod_{I=1}^{n} \prod_{g=1}^{m} (P_{gI})^{S_{gI}} \to \max ,$$

ahol $S_{gI} = 1$, ha a $g$ alternatívát az $I$ személy kiválasztotta, egyébként $S_{gI} = 0$. A paraméterek referencia-függők, a valószínűségek viszont nem. A valószínűséget a

$$P_{CI} = \frac{1}{\sum_{g=1}^{m} e^{(\alpha_g - \alpha_C) + (\beta_g - \beta_C) x_I}}$$

formában írva látszik, hogy a különféle alternatívák kiválasztási valószínűségei közötti különbség adott $x$ mellett csak a paraméterek alternatíva-specifikus jellegéből származik.

Illusztratív példánkban $I = 1, 2, \ldots, 21$ személynek az utazásuk módját illető választásait az életkorukkal magyarázzuk, tengelymetszet szerepeltetése mellett. A lehetséges három mód: *autó* ($A$), *repülő* ($R$) és *vonat* ($V$). A három mód közül egyet és csak egyet választ mindenki. A három mód mellett változónként rendre 2 koefficienst (két tengelymetszetet és két életkormeredekséget) becslünk úgy, hogy a *vonat* ($V$) a referencia alternatíva. A

koefficiensek ML becsléseit az *1. tábla*, az adatokat és az eredményeket pedig a *2. tábla* közli. A táblában a „$C$" megnevezésű oszlop az illető választási döntését közli, az $S_A$, $S_R$ és $S_V$ indikátor jellegű oszlopok pedig a döntés statusa szerint veszik fel az 1, illetve a 0 értéket.

Például az $I = 1$ egyén, aki egyébként 32 éves, a repülős utat választotta, ezért

$$\text{odds}(R/V \mid 32) = e^{2.7212 - 0.05 \cdot 32} = 3.068 \,,$$

majd a repülős út kiválasztásának a valószínűsége

$$P_{R1} = \frac{3.068}{2.168 + 3.068 + 1} = 0.492 \,.$$

Ez a valószínűség bárkit, aki 32 éves, egyöntetűen jellemez! Az így számított 21 darab $P_C$ valószínűség $L = 0.492 \cdot 0.483 \cdot \ldots \cdot 0.421$ szorzata a 2. táblában az 1. tábla koefficiensei mellett maximális, a megfelelő $-2 \ln L$ „*goodness-of-fit*" statisztika értéke pedig 42.18.[4]

| Változó | Koefficiens | | |
|---|---|---|---|
| | Autó | Repülő | Vonat |
| Tengelymetszet | 3.0449 | 2.7212 | 0 |
| Életkor | -0.0710 | -0.0500 | 0 |

*1. táblázat.* ML becslés az életkor-koefficiensekre

| Megfigyelés (I) | Életkor | $C$ | $S_A$ | $S_R$ | $S_V$ | $odds_A$ | $odds_R$ | $odds_V$ | $P_C$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 32 | R | 0 | 1 | 0 | 2.168 | 3.068 | 1 | 0.492 |
| 2 | 13 | A | 1 | 0 | 0 | 8.351 | 7.934 | 1 | 0.483 |
| 3 | 41 | V | 0 | 0 | 1 | 1.145 | 1.956 | 1 | 0.244 |
| 4 | 41 | V | 0 | 0 | 1 | 1.145 | 1.956 | 1 | 0.244 |
| 5 | 47 | A | 1 | 0 | 0 | 0.748 | 1.449 | 1 | 0.234 |
| 6 | 24 | R | 0 | 1 | 0 | 3.826 | 4.577 | 1 | 0.487 |
| 7 | 27 | A | 1 | 0 | 0 | 3.092 | 3.940 | 1 | 0.385 |
| 8 | 21 | R | 0 | 1 | 0 | 4.733 | 5.318 | 1 | 0.481 |
| 9 | 23 | A | 1 | 0 | 0 | 4.107 | 4.812 | 1 | 0.414 |
| 10 | 30 | R | 0 | 1 | 0 | 2.499 | 3.391 | 1 | 0.492 |
| 11 | 58 | R | 0 | 1 | 0 | 0.343 | 0.836 | 1 | 0.384 |
| 12 | 36 | V | 0 | 0 | 1 | 1.633 | 2.512 | 1 | 0.194 |
| 13 | 43 | A | 1 | 0 | 0 | 0.993 | 1.770 | 1 | 0.264 |
| 14 | 33 | R | 0 | 1 | 0 | 2.020 | 2.919 | 1 | 0.491 |
| 15 | 30 | R | 0 | 1 | 0 | 2.499 | 3.391 | 1 | 0.492 |
| 16 | 28 | R | 0 | 1 | 0 | 2.880 | 3.748 | 1 | 0.491 |
| 17 | 44 | R | 0 | 1 | 0 | 0.925 | 1.684 | 1 | 0.467 |
| 18 | 37 | V | 0 | 0 | 1 | 1.521 | 2.390 | 1 | 0.204 |
| 19 | 45 | A | 1 | 0 | 0 | 0.862 | 1.602 | 1 | 0.249 |
| 20 | 35 | R | 0 | 1 | 0 | 1.753 | 2.641 | 1 | 0.490 |
| 21 | 22 | A | 1 | 0 | 0 | 4.409 | 5.059 | 1 | 0.421 |

*2. táblázat.* Utazási mód választása adott életkor (év) mellett

[4] A maximálást a paraméterek tekintetében —az 1–5. táblák esetén— a MS Office 2003 Excel-Solver moduljával végeztük el. Ezzel pontbecslést nyertünk a paraméterekre.

A fenti modell becslése statisztikai szoftverekben standard módon, közvetlenül elérhető úgy, hogy 21 megfigyelés (case) mellett az *Életkor* a kovariáns és $C$ az eredményváltozó.

Ha több, rendre $x_1, x_2, \ldots, x_k$ magyarázó változót tartalmaz a modell, akkor a regressziós paraméterek köre is megfelelően bővül, de a modell lényegileg változatlan marad. Például három magyarázó változót használva (*Életkor, Jövedelem, Nem*) három alternatívához a két tengelymetszet mellett még $3 * 2 = 6$ meredekséget is becsülnünk kell. Ha pedig olyan nominális magyarázó változóval bővítjük a modellt, melynek kettőnél több kimenete van (lakóhely szerint Budapest, Többi_város, Község), akkor az indikátor változók száma kettővel (B, Tv), a paraméterek száma pedig $2 * 2 = 4$-gyel emelkedik. Mint látható, a PL modell nem takarékos a paraméterekkel, és a paraméterek értelmezése a mindenkori referencia alternatíva viszonylatát igényli.

# 3 A feltételes logit modell esete

A feltételes (conditional) logit modell (CL) akkor alkalmazandó, ha a döntés kimenetét magyarázó változó *alternatíva-specifikus*, annak valamely tulajdonságát írja le. Jelölje $Z$ az alternatívák egy jellemzőjét: példánkban az utazás időigényét. Megfigyelésünk most nem az individuumra, hanem az összes előforduló $Z_{gI}$ ($I = 1, 2, \ldots, n$; $g = 1, 2, \ldots, m_I$) utazási időre (órában mérve) irányul. A megfigyelt eseteket az individuumok $n$ számú rétegre bontják, és adott rétegen belül egy alternatíva kiválasztásra kerül, a többi nem.

A $21 * 3 = 63$ megfigyelést (esetet) 21 rétegre bontva a *3. tábla* tartalmazza. Az egyes alternatívák időigényét adott individuum esetén az $AI$, $RI$ és $VI$ oszlopok azonosítják, de magyarázó változónk csak egy van, az *utazás időigénye*. Valamennyi *esetet* szemlélve a döntés kimenetét binárisan kódoljuk: $S_{gI} = 1$, ha a $g$ alternatíva kiválasztásra került az $I$ rétegben, és $S_{gI} = 0$, ha nem. Az eredményváltozó megfelelő rétegzett értéke tehát: $S_{gI}$. A kiválasztás előrejelzése így egy *rétegzett, dichotom* logit modell alkalmazására vezetett. E modell paramétere alternatíva-független, globális, minden megfigyelésre egyformán érvényes. E paraméterek becslése a CL modell speciális alkalmazásával valósítható meg.

A CL modell lényegét segít megvilágítani, ha előbb külön az $I = 1$ individuumot (réteget) tekintjük, aki a repülőt választotta, tehát esetében a háromelemű bináris döntési szekvencia: $\mathbf{d}_R = [0, 1, 0]$. Ennek likelihoodja a dichotom logit modell alapján:

$$L_{RI} = \frac{1}{1 + e^{\alpha + \theta Z_{AI}}} \cdot \frac{e^{\alpha + \theta Z_{RI}}}{1 + e^{\alpha + \theta Z_{RI}}} \cdot \frac{1}{1 + e^{\alpha + \theta Z_{VI}}} ,$$

ahol $\alpha$ és $\theta$ *globális* paraméterek. A döntéshozó azonban választhatott volna másképpen is. Ragaszkodva ahhoz, hogy csak egy alternatívát választhat, a további lehetőségei rendre a $\mathbf{d}_A = [1, 0, 0]$ és a $\mathbf{d}_V = [0, 0, 1]$ szekvenciák,

melyek likelihoodjai rendre:

$$L_{AI} = \frac{e^{\alpha+\theta Z_{AI}}}{1+e^{\alpha+\theta Z_{AI}}} \cdot \frac{1}{1+e^{\alpha+\theta Z_{RI}}} \cdot \frac{1}{1+e^{\alpha+\theta Z_{VI}}} ,$$

$$L_{VI} = \frac{1}{1+e^{\alpha+\theta Z_{AI}}} \cdot \frac{1}{1+e^{\alpha+\theta Z_{RI}}} \cdot \frac{e^{\alpha+\theta Z_{VI}}}{1+e^{\alpha+\theta Z_{VI}}} .$$

Ezek birtokában a $C \in \{A, R, V\}$ alternatíva kiválasztásának a feltételes valószínűsége az $I$ egyén esetében a megfelelő likelihood statisztikai *megoszlása* a három likelihood összegében, általában pedig:

$$P_{CI} = \frac{L_{CI}}{\sum_{g=1}^{m_I} L_{gI}} = \frac{e^{\theta Z_{CI}}}{\sum_{g=1}^{m_I} e^{\theta Z_{gI}}} = \frac{1}{\sum_{g=1}^{m_I} e^{\theta(Z_{gI}-Z_{CI})}} .$$

Vegyük észre, hogy a likelihoodok közös nevezője és a globális tengelymetszet eliminálódik a valószínűségből, ezért szerepeltetésük fölösleges.[5] A magyarázó változó egységnyi abszolút növekményének a kiválasztási valószínűségre gyakorolt hatása a magyarázó $Z$ változó alternatívák közötti ingadozásától függ, és konstans.

Végül a $\theta$ paraméter tekintetében maximálandó likelihood a rétegen belüli kiválasztási valószínűségek szorzata:

$$L = \prod_{I=1}^{n} \prod_{g=1}^{m_I} (P_{gI})^{S_{gI}} \to \max .$$

Az egyetlen változónkhoz tartozó $\theta$ paraméter ML becslése $\hat{\theta} = -0.26549$. Így az $I = 1$ egyén esetén a repülős út kiválasztásának a valószínűsége:

$$P_{RI} = \frac{0.303}{0.07 + 0.303 + 0.062} = 0.697 .$$

Az ily módon kalkulált 21 darab $P_C$ valószínűség $L$ szorzata a *3. táblában* a fenti koefficiens mellett maximális, és a $-2 \ln L$ statisztika értéke 33.629.

Ha több, rendre $Z_1, Z_2, \ldots, Z_q$ magyarázó változót tartalmaz a modell, akkor a regressziós paraméterek köre is megfelelően bővül.

[5] A feltételes logit $P_{CI}$ valószínűségének a nevezője azért tartalmaz annyi összeadandót, ahány alternatíva van, mert az individuum csak egy alternatívát választhat ki, így az egy darab 1 összes permutációinak a száma megegyezik az alternatívák számával. Ha háromnál több alternatíva közül egynél többet is választhatunk, például kettőt, akkor az összes olyan permutációk száma melyek a szekvenciában két helyen tartalmaznak 1 értéket, már megsokszorozódik.

| Réteg ($I$) | AI | RI | VI | $S_A$ | $S_R$ | $S_V$ | $e^{\theta AI}$ | $e^{\theta RI}$ | $e^{\theta VI}$ | $P_C$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 4.5 | 10.5 | 0 | 1 | 0 | 0.070 | 0.303 | 0.062 | 0.697 |
| 2 | 5.5 | 4 | 7.5 | 1 | 0 | 0 | 0.232 | 0.346 | 0.137 | 0.325 |
| 3 | 4.5 | 6 | 5.5 | 0 | 0 | 1 | 0.303 | 0.203 | 0.232 | 0.314 |
| 4 | 3.5 | 2 | 5 | 0 | 0 | 1 | 0.395 | 0.588 | 0.265 | 0.212 |
| 5 | 1.5 | 4.5 | 4 | 1 | 0 | 0 | 0.671 | 0.303 | 0.346 | 0.509 |
| 6 | 10.5 | 3 | 10.5 | 0 | 1 | 0 | 0.062 | 0.451 | 0.062 | 0.786 |
| 7 | 7 | 3 | 9 | 1 | 0 | 0 | 0.156 | 0.451 | 0.092 | 0.223 |
| 8 | 9 | 3.5 | 9 | 0 | 1 | 0 | 0.092 | 0.395 | 0.092 | 0.683 |
| 9 | 4 | 5 | 5.5 | 1 | 0 | 0 | 0.346 | 0.265 | 0.232 | 0.410 |
| 10 | 22 | 4.5 | 22.5 | 0 | 1 | 0 | 0.003 | 0.303 | 0.003 | 0.982 |
| 11 | 7.5 | 5.5 | 10 | 0 | 1 | 0 | 0.137 | 0.232 | 0.070 | 0.529 |
| 12 | 11.5 | 3.5 | 11.5 | 0 | 0 | 1 | 0.047 | 0.395 | 0.047 | 0.096 |
| 13 | 3.5 | 4.5 | 4.5 | 1 | 0 | 0 | 0.395 | 0.303 | 0.303 | 0.395 |
| 14 | 12 | 3 | 11 | 0 | 1 | 0 | 0.041 | 0.451 | 0.054 | 0.826 |
| 15 | 18 | 5.5 | 20 | 0 | 1 | 0 | 0.008 | 0.232 | 0.005 | 0.946 |
| 16 | 23 | 5.5 | 21.5 | 0 | 1 | 0 | 0.002 | 0.232 | 0.003 | 0.977 |
| 17 | 4 | 3 | 4.5 | 0 | 1 | 0 | 0.346 | 0.451 | 0.303 | 0.410 |
| 18 | 5 | 2.5 | 7 | 0 | 0 | 1 | 0.265 | 0.515 | 0.156 | 0.167 |
| 19 | 3.5 | 2 | 7 | 1 | 0 | 0 | 0.395 | 0.588 | 0.156 | 0.347 |
| 20 | 12.5 | 3.5 | 15.5 | 0 | 1 | 0 | 0.036 | 0.395 | 0.016 | 0.883 |
| 21 | 1.5 | 4 | 2 | 1 | 0 | 0 | 0.671 | 0.346 | 0.588 | 0.418 |

*3. táblázat.* Utazási mód választása az utazási idő (óra) függvényében

# 4 A „vegyes” modell alkalmazása

A valósághű alkalmazás mind a választó individuum, mind a választandó alternatíva jegyeit figyelembe veszi. Ekkor a $C$ alternatíva $I$ egyén által való kiválasztásának a valószínűsége:

$$P_{CI} = \frac{e^{\alpha_C + \beta_C x_I + \theta Z_{CI}}}{\sum_{g=1}^{m} e^{\alpha_g + \beta_g x_I + \theta Z_{gI}}} \quad | \, \alpha_m = \beta_m = 0 \, .$$

A ML becsléssel nyert koefficienseket a *4. tábla* közli, melyekre az *5. táblában* foglalt 21 db $P_C$ valószínűség szorzata maximális: a $-2 \ln L$ statisztika értéke 27.46433.

| Változó | Koefficiens | | | |
|---|---|---|---|---|
| | Globális | Autó | Repülő | Vonat |
| Tengelymetszet | | 2.5007 | -2.7792 | 0 |
| Életkor | | -0.07830 | 0.0169 | 0 |
| Utazási idő | -0.6085 | | | |

*4. táblázat.* A vegyes modell ML koefficiensei

| $I$ | $AI$ | $RI$ | $VI$ | $Kor$ | $C$ | $2.5 + -.078Kor + -.608AI$ | $-2.779 + .017Kor + -.608RI$ | $-.608VI$ | $P_C$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 4.5 | 10.5 | 32 | R | −6.089 | −4.975 | −6.389 | 0.636 |
| 2 | 5.5 | 4 | 7.5 | 13 | A | −1.864 | −4.993 | −4.564 | 0.900 |
| 3 | 4.5 | 6 | 5.5 | 41 | V | −3.446 | −5.735 | −3.347 | 0.501 |
| 4 | 3.5 | 2 | 5 | 41 | V | −2.838 | −3.301 | −3.042 | 0.333 |
| 5 | 1.5 | 4.5 | 4 | 47 | A | −2.090 | −4.721 | −2.434 | 0.561 |
| 6 | 10.5 | 3 | 10.5 | 24 | R | −5.766 | −4.197 | −6.389 | 0.757 |
| 7 | 7 | 3 | 9 | 27 | A | −3.871 | −4.147 | −5.476 | 0.510 |
| 8 | 9 | 3.5 | 9 | 21 | R | −4.619 | −4.553 | −5.476 | 0.428 |
| 9 | 4 | 5 | 5.5 | 23 | A | −1.733 | −5.431 | −3.347 | 0.817 |
| 10 | 22 | 4.5 | 22.5 | 30 | R | −13.233 | −5.009 | −13.691 | 0.999 |
| 11 | 7.5 | 5.5 | 10 | 58 | R | −6.602 | −5.143 | −6.085 | 0.616 |
| 12 | 11.5 | 3.5 | 11.5 | 36 | V | −7.314 | −4.299 | −6.997 | 0.060 |
| 13 | 3.5 | 4.5 | 4.5 | 43 | A | −2.994 | −4.788 | −2.738 | 0.407 |
| 14 | 12 | 3 | 11 | 33 | R | −7.384 | −4.045 | −6.693 | 0.904 |
| 15 | 18 | 5.5 | 20 | 30 | R | −10.799 | −5.618 | −12.169 | 0.993 |
| 16 | 23 | 5.5 | 21.5 | 28 | R | −13.685 | −5.652 | −13.082 | 0.999 |
| 17 | 4 | 3 | 4.5 | 44 | R | −3.377 | −3.858 | −2.738 | 0.176 |
| 18 | 5 | 2.5 | 7 | 37 | V | −3.437 | −3.673 | −4.259 | 0.197 |
| 19 | 3.5 | 2 | 7 | 45 | A | −3.151 | −3.234 | −4.259 | 0.444 |
| 20 | 12.5 | 3.5 | 15.5 | 35 | R | −7.844 | −4.316 | −9.431 | 0.966 |
| 21 | 1.5 | 4 | 2 | 22 | A | −0.134 | −4.840 | −1.217 | 0.742 |

*5. táblázat.* Utazási mód választása az életkor (év) és az utazási idő (óra) függvényében

A vegyes modellben annak a valószínűsége, hogy az $I = 1$ személy a repülős utat választja:

$$P_{R1} = \frac{e^{-4.975}}{e^{-6.089} + e^{-4.975} + e^{-6.389}} = 0.636 \, .$$

Mint látható, a vegyes modell individuális változójának paramétere alternatíva-specifikus, míg az alternatíva-specifikus változó paramétere globális. Ez nehézséget okoz akkor, ha *szimultán* becslésükre (az alkalmazott statisztikai szoftver adottsága miatt) a polichotom logit módszer és a feltételes logit módszer csak szeparáltan hívható fel. Kézenfekvő megoldás az individuális változót is globalizálni.

# 5 A vegyes modell globális paraméterezése

Tekintsük az $i = 1, 2, \ldots, nm$ egyedi választási (utazási) lehetőségeket, melyeket az individuumok $n$ rétegbe sorolnak. Definiáljuk az $X$ *globális* változó értékeit az $X_i$ $(i = 1, 2, \ldots, nm)$ módon, ahol $i$ egy $(g, I)$ párosítást képvisel. Azonosítsa továbbá a $D_g$ globális *indikátor* változó 1 értékkel a $g$ alternatívát, értéke egyébként 0. Így, az indikátor változók felhasználásával $X$ hatása a kiválasztásra a $\gamma_0 + \gamma_1 X_i$ modell szerint alakul, ahol a $\gamma$ globális paraméterek alternatívafüggők, az alábbiaknak megfelelően:

$$\gamma_0 = \sum_{g=1}^{m} \alpha_g D_g \, , \quad \gamma_1 = \sum_{g=1}^{m} \beta_g D_g \, , \quad (\alpha_m = \beta_m = 0) \, .$$

Így a globális $X$ változó lineáris hatása:

$$\sum_{g=1}^{m} \alpha_g D_g + \sum_{g=1}^{m} \beta_g (D_g X_i) \ .$$

Alkossák most a $\mathbf{Z}$ változók körét egyfelől az eredetileg is $Z$ jellegű változó, másfelől az alternatívát azonosító $D_g$ változók, végül ezen indikátor változóknak az $X$ változóval vett $D_g * X = D_g X$ *interakciói*. A $C$ utazási mód kiválasztásának a valószínűsége az $I$ individuum által (az életkort és az utazás idejét egyidejűleg figyelembe véve):

$$P_{CI} = \frac{e^{\alpha_C D_C + \beta_C (D_C X_{CI}) + \theta Z_{CI}}}{\sum_{g=1}^{m} e^{\alpha_g D_g + \beta_g (D_g X_{gI}) + \theta Z_{gI}}} \quad (\alpha_m = \beta_m = 0) \ ,$$

ahol a paraméterek becslése a feltételes likelihood maximálását igényli.

A fenti módon definiált, példabeni adatainkat a *6. tábla* írja le. E struktúrában adott individuum választási halmaza egy önálló réteget (strata) alkot, melyen belül mindegyik alternatíva egy önálló megfigyelést (sort) igényel. Az adatállományban a sorok száma $21 * 3 = 63$, a rétegek száma 21, és mindegyik réteg 3 alternatívát tartalmaz. Az $S$ indikátor változó azt jelzi, hogy az alternatíva kiválasztásra került vagy sem. A $D_A$ és $D_R$ indikátor változók rendre az ,,autós" és a ,,repülős" utat azonosítják, miközben a ,,vonatutazás" a *referencia* alternatíva. (A $t$ oszlop tartalma a következő fejezetben kerül definiálásra.)

Ha a paraméterbecsléshez feltételes maximum likelihood program nem áll rendelkezésre, akkor a probléma *túlélési* modellként való megfogalmazása nyújt megfelelő eredményt. Ennek során minden kiválasztást mint bekövetkezett eseményt, a ki nem választásokat pedig mint később bekövetkezendő eseményeket kezeljük, majd az ,,eseményig" tartó időtartam alakulását modellezzük magyarázó változók ismerete mellett.

Ennek egyik eszköze a *Cox*-regresszió, mely speciális körülmények között a CL modell megoldását nyújtja (Kuhfeld (2003)).

| $I$ | Mód | Idő | Kor | $S$ | $t$ | $D_A$ | $D_R$ | $D_A * Kor$ | $D_R * Kor$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | A | 10.0 | 32 | 0 | 2 | 1 | 0 | 32 | 0 |
| 1 | R | 4.5 | 32 | 1 | 1 | 0 | 1 | 0 | 32 |
| 1 | V | 10.5 | 32 | 0 | 2 | 0 | 0 | 0 | 0 |
| 2 | A | 5.5 | 13 | 1 | 1 | 1 | 0 | 13 | 0 |
| 2 | R | 4.0 | 13 | 0 | 2 | 0 | 1 | 0 | 13 |
| 2 | V | 7.5 | 13 | 0 | 2 | 0 | 0 | 0 | 0 |
| 3 | A | 4.5 | 41 | 0 | 2 | 1 | 0 | 41 | 0 |
| 3 | R | 6.0 | 41 | 0 | 2 | 0 | 1 | 0 | 41 |
| 3 | V | 5.5 | 41 | 1 | 1 | 0 | 0 | 0 | 0 |
| 4 | A | 3.5 | 41 | 0 | 2 | 1 | 0 | 41 | 0 |
| 4 | R | 2.0 | 41 | 0 | 2 | 0 | 1 | 0 | 41 |
| 4 | V | 5.0 | 41 | 1 | 1 | 0 | 0 | 0 | 0 |
| 5 | A | 1.5 | 47 | 1 | 1 | 1 | 0 | 47 | 0 |
| 5 | R | 4.5 | 47 | 0 | 2 | 0 | 1 | 0 | 47 |
| 5 | V | 4.0 | 47 | 0 | 2 | 0 | 0 | 0 | 0 |

*6. táblázat.* Vegyes modell interakciókkal

| $I$ | Mód | Idő | Kor | $S$ | $t$ | $D_A$ | $D_R$ | $D_A * Kor$ | $D_R * Kor$ |
|---|---|---|---|---|---|---|---|---|---|
| 6 | A | 10.5 | 24 | 0 | 2 | 1 | 0 | 24 | 0 |
| 6 | R | 3.0 | 24 | 1 | 1 | 0 | 1 | 0 | 24 |
| 6 | V | 10.5 | 24 | 0 | 2 | 0 | 0 | 0 | 0 |
| 7 | A | 7.0 | 27 | 1 | 1 | 1 | 0 | 27 | 0 |
| 7 | R | 3.0 | 27 | 0 | 2 | 0 | 1 | 0 | 27 |
| 7 | V | 9.0 | 27 | 0 | 2 | 0 | 0 | 0 | 0 |
| 8 | A | 9.0 | 21 | 0 | 2 | 1 | 0 | 21 | 0 |
| 8 | R | 3.5 | 21 | 1 | 1 | 0 | 1 | 0 | 21 |
| 8 | V | 9.0 | 21 | 0 | 2 | 0 | 0 | 0 | 0 |
| 9 | A | 4.0 | 23 | 1 | 1 | 1 | 0 | 23 | 0 |
| 9 | R | 5.0 | 23 | 0 | 2 | 0 | 1 | 0 | 23 |
| 9 | V | 5.5 | 23 | 0 | 2 | 0 | 0 | 0 | 0 |
| 10 | A | 22.0 | 30 | 0 | 2 | 1 | 0 | 30 | 0 |
| 10 | R | 4.5 | 30 | 1 | 1 | 0 | 1 | 0 | 30 |
| 10 | V | 22.5 | 30 | 0 | 2 | 0 | 0 | 0 | 0 |
| 11 | A | 7.5 | 58 | 0 | 2 | 1 | 0 | 58 | 0 |
| 11 | R | 5.5 | 58 | 1 | 1 | 0 | 1 | 0 | 58 |
| 11 | V | 10.0 | 58 | 0 | 2 | 0 | 0 | 0 | 0 |
| 12 | A | 11.5 | 36 | 0 | 2 | 1 | 0 | 36 | 0 |
| 12 | R | 3.5 | 36 | 0 | 2 | 0 | 1 | 0 | 36 |
| 12 | V | 11.5 | 36 | 1 | 1 | 0 | 0 | 0 | 0 |
| 13 | A | 3.5 | 43 | 1 | 1 | 1 | 0 | 43 | 0 |
| 13 | R | 4.5 | 43 | 0 | 2 | 0 | 1 | 0 | 43 |
| 13 | V | 4.5 | 43 | 0 | 2 | 0 | 0 | 0 | 0 |
| 14 | A | 12.0 | 33 | 0 | 2 | 1 | 0 | 33 | 0 |
| 14 | R | 3.0 | 33 | 1 | 1 | 0 | 1 | 0 | 33 |
| 14 | V | 11.0 | 33 | 0 | 2 | 0 | 0 | 0 | 0 |
| 15 | A | 18.0 | 30 | 0 | 2 | 1 | 0 | 30 | 0 |
| 15 | R | 5.5 | 30 | 1 | 1 | 0 | 1 | 0 | 30 |
| 15 | V | 20.0 | 30 | 0 | 2 | 0 | 0 | 0 | 0 |
| 16 | A | 23.0 | 28 | 0 | 2 | 1 | 0 | 28 | 0 |
| 16 | R | 5.5 | 28 | 1 | 1 | 0 | 1 | 0 | 28 |
| 16 | V | 21.5 | 28 | 0 | 2 | 0 | 0 | 0 | 0 |
| 17 | A | 4.0 | 44 | 0 | 2 | 1 | 0 | 44 | 0 |
| 17 | R | 3.0 | 44 | 1 | 1 | 0 | 1 | 0 | 44 |
| 17 | V | 4.5 | 44 | 0 | 2 | 0 | 0 | 0 | 0 |
| 18 | A | 5.0 | 37 | 0 | 2 | 1 | 0 | 37 | 0 |
| 18 | R | 2.5 | 37 | 0 | 2 | 0 | 1 | 0 | 37 |
| 18 | V | 7.0 | 37 | 1 | 1 | 0 | 0 | 0 | 0 |
| 19 | A | 3.5 | 45 | 1 | 1 | 1 | 0 | 45 | 0 |
| 19 | R | 2.0 | 45 | 0 | 2 | 0 | 1 | 0 | 45 |
| 19 | V | 7.0 | 45 | 0 | 2 | 0 | 0 | 0 | 0 |
| 20 | A | 12.5 | 35 | 0 | 2 | 1 | 0 | 35 | 0 |
| 20 | R | 3.5 | 35 | 1 | 1 | 0 | 1 | 0 | 35 |
| 20 | V | 15.5 | 35 | 0 | 2 | 0 | 0 | 0 | 0 |
| 21 | A | 1.5 | 22 | 1 | 1 | 1 | 0 | 22 | 0 |
| 21 | R | 4.0 | 22 | 0 | 2 | 0 | 1 | 0 | 22 |
| 21 | V | 2.0 | 22 | 0 | 2 | 0 | 0 | 0 | 0 |

*6. táblázat.* Vegyes modell interakciókkal (folyt.)

# 6 A Cox-regresszió: „proportional hazards”

Jelölje $t$ a vizsgált „*esemény*” bekövetkezéséig a megfigyelés (folyamat) kezdetétől eltelt idő hosszát: „*event time*”. E periódus változó időtartamát a

modell szerint a $Z_1$, $Z_2$, ... magyarázó változók szintjei indokolják, és $t_j$ a megfigyelt időtartamok *növekvő* rangsorában a $j$-edik, miközben $f_j$ annak a gyakorisága, hogy $t_j$ eltelt idő mellett a vizsgált eseményt hányszor észleltük:[6]

$$t_{1(f_1)} < t_{2(f_2)} < \ldots < t_{j(f_j)} < \ldots t_{k(f_k)} \ .$$

Ha egy individuum —akinél a folyamat már elindult, de— valami ok folytán kikerül a megfigyelési körből az *esemény bekövetkezése nélkül*, akkor az illető megfigyelést *cenzorált* (censored) esetként kezeljük. Jelölje továbbá $R_j$ mindazon indexek által alkotott kockázati csoportot, akik közvetlen a $t_j$ időt megelőzőleg ki vannak téve az esemény kockázatának. E kockázati körben az ,,event time" *legalább* $t_j$, és a $t_j$ mellett cenzorált esetek tagjai e kockázati csoportnak. Ekkor annak feltételes valószínűsége, hogy valamely individuum megéli a $t_j$ időt, de utána az esemény rögtön bekövetkezik, nem más, mint a ,,*hazard-ratio*" megoszlása:

$$P_Z = \frac{e^{\beta^T Z}}{\sum_{l \in R_j} e^{\beta^T Z_l}} \ .$$

E valószínűségek szorzata valamennyi $t$ időre (súlyozottan felírva) a Breslow-féle likelihood függvényt adja:

$$L(\beta) = \prod_{j=1}^{k} \frac{e^{\beta^T Z_j^+}}{\left(\sum_{l \in R_j} e^{\beta^T Z_l}\right)^{f_j}} \to \max \ ,$$

ahol $Z_j^+$ a megfelelő magyarázó változó összegzését jelöli mindazokra, akiknél az esemény a $t_j$ időpontban bekövetkezett. (A súlyozatlan eset, mikor $f_j = 1$ minden $j$-re, speciálisan a Cox-féle parciális likelihood függvényt eredményezi.) Ha a minta $I = 1, 2, \ldots, n$ rétegre van bontva, akkor a Breslow-likelihood egyszerűen a rétegen belüli likelihoodok szorzata:

$$L(\beta) = \prod_{I=1}^{n} L_I(\beta) \ .$$

Ahhoz, hogy a Breslow-likelihood a feltételes logit likelihoodjával ekvivalens legyen, az alábbiak szükségesek:

1. A megfigyeléseket az individuumok szerinti rétegekre (strata) bontjuk,
2. a kiválasztott $C$ alternatívához a status változóban $S = 1$ (event) értéket, a ki nem választott alternatívákhoz pedig az $S = 0$ (censored) értéket rendeljük,
3. a kiválasztott $C$ alternatívához mindig $t = 1$ ,,event time", a ki nem választott alternatívákhoz pedig egy nagyobb (későbbi), de egyöntetűen $t = 2$ ,,censored time" értéket rendelünk,

[6] A túlélési modell néhány alapfogalmát lásd a Függelékben!

4. mivel a diszkrét kiválasztási modellben a „$t$" *változó* adott értéke szükségszerűen többször fordul elő, ezért ha e kötések (ties) kezelésére az alkalmazott szoftverben opcionálisan más típus is választható (lásd SAS), akkor kifejezetten a Breslow-likelihood választandó.

A 6. tábla adataira a Cox-regressziót alkalmazva visszakapjuk a korábban már megismert (Excel-Solver) megoldásokat, az alábbiak szerint:

1. rétegképző „strata" változó: „*Individuum*",
2. a „status" változó: $S$,
3. az „event time" változó: $t$,
4. a kovariánsok: *Utazási idő*, $D_A$, $D_R$, $D_A * Kor$, $D_R * Kor$.

A ($B$) pontbecslések mellett aszimptotikus standard hibákat ($SE$), parciális Wald-statisztikákat, szignifikancia-értékeket, $\exp(B)$ „*hazard-ratio*" értékeket és 95%-os konfidencia intervallumokat is nyerünk. Az SPSS programmal kapott eredményeket a *7. tábla* közli.

Eszerint 5 százalékos szignifikancia szinten csak az utazás időtartama hat szignifikánsan a választásra. Továbbmenve, ha az utazás 1 órával tovább tart, akkor a kérdéses utazási mód kiválasztásának az esélye $100 \cdot (1 - 0.544) = 45.6$ százalékkal csökken. A többi paraméter tesztelése és az $\exp(B)$ „*hazard-ratio*" értelmezése analóg.

Az előzőekben modellenként rendre közöltük a Likelihood Ratio típusú *goodness-of-fit* statisztikák $-2 \ln L$ értékeit. A háromféle modell úgy ítélendő meg, hogy a tökéletesen illeszkedő *szaturált modell* esetén $-2 \ln L = 0$, míg a kovariánst nem tartalmazó „intercept only" ún. *null-modell* esetén $-2 \ln L = 46.142$. E határok között az egyes változók lépésenkénti szelektálására is lehetőség nyílik, melynek eredményeit a *8. tábla* közli.

| Változó | $B$ | $SE$ | Wald | *df* | $p$-value | $\exp(B)$ | Lower | Upper |
|---|---|---|---|---|---|---|---|---|
| Utazási idő | -.608 | .271 | 5.031 | 1 | .025 | .544 | .320 | .926 |
| $D_A$ | 2.501 | 2.396 | 1.089 | 1 | .297 | 12.191 | .111 | 1334.724 |
| $D_R$ | -2.779 | 3.529 | .620 | 1 | .431 | .062 | .000 | 62.686 |
| $D_A * Kor$ | -.078 | .063 | 1.527 | 1 | .217 | .925 | .817 | 1.047 |
| $D_R * Kor$ | .017 | .074 | .052 | 1 | .820 | 1.017 | .879 | 1.177 |

*7. táblázat.* A vegyes modell paraméterbecslése a Cox-regresszióból

| Bevont változó | $-2 \ln L$ | $Chi2$ | *df* | $p$-value | $Chi2$_változás | *df* | $p$-value |
|---|---|---|---|---|---|---|---|
| 1.: Utazási idő | 33.629 | 11.988 | 1 | .001 | 12.513 | 1 | 0.000 |
| 2.: $D_R$ | 30.284 | 13.522 | 2 | .001 | 3.345 | 1 | 0.067 |
| 3.: $D_R * Kor$ | 29.266 | 13.940 | 3 | .003 | 1.018 | 1 | 0.313 |
| 4.: $D_A * Kor$ | 28.739 | 13.966 | 4 | .007 | 0.527 | 1 | 0.468 |
| 5.: $D_A$ | 27.464 | 15.361 | 5 | .009 | 1.274 | 1 | 0.259 |

*8. táblázat.* A likelihood-arány javulása változóról változóra

Első lépésben csak az *Utazási idő*, az utolsó lépésben pedig mind az öt magyarázó változó a modellben szerepel. A *null*-modelltől való eltávolodást mérő Chi2 statisztika még a legbővebb modellt is szignifikánsnak ítéli 1%-os szinten, bár a *df* szabadsági fok a modell komplexitásának növekedésével gyorsabban nőtt, mint ahogy a $-2\ln L$ célfüggvény csökkent. A Chi2 lépésenkénti változását tesztelve látszik, hogy az utolsó három lépésben bevont tényező modellből való kihagyása megfontolandó.

# 7 Függetlenség az irreleváns alternatíváktól

Az eddigi modellek mindegyike azon a feltevésen alapult, hogy az alternatívák választása független egymástól: „Independence from Irrelevant Alternatives" (IIA). Ez alatt az értendő, hogy adott megfigyelésre bármely két alternatíva kiválasztási valószínűségének az egymáshoz való OR (odds-ratio) aránya független bármely más alternatívától. E feltevés lehet helytálló, lehet irreális, viszont fenntartása vagy elvetése statisztikai tesztet igényel.

Esetünkben az utazási *módok* és az utazási *idők* kölcsönhatásai (interakciói) valamint az utazási *módok egymás közti* kapcsolatainak az utazási időre gyakorolt hatása vizsgálható.

Az alternatívák és időigényeik interakcióit megfogalmazó modell

$$\sum_{g=1}^{m-1} \alpha_g D_g + \sum_{g=1}^{m} \theta_g D_g Z \ ,$$

melyet tovább bővítve alternatívaközi interakciók hozzáadásával

$$\sum_{g=1}^{m-1} \alpha_g D_g + \sum_{g=1}^{m} \theta_g D_g Z + \sum_{g=2}^{m} \delta_g D_g Z_{g-1} + \delta_1 D_1 Z_m$$

adódik. Láthatóan az alternatívák közötti kapcsolatok tesztelését most a szomszédos alternatívák vizsgálatára egyszerűsítettük. A két egymásba ágyazott modell közötti választás a paraméterek egy csoportjára vonatkozó hipotézis tesztelését igényli:

$$\delta_1 = \delta_2 = \ldots = \delta_m = 0 \ .$$

Az újonnan bevezetett magyarázó változókat —a korábban definiált változókkal együtt— a *9. tábla* tartalmazza. Például $D_R * AI$ a repülős alternatíva indikátor változójának és az autóval való utazási időnek a szorzata (repülőn utazva tart addig az út, mint egyébként autón). Az eredményeket a tágabb, meg nem szorított modellre a *10. tábla*, a szűkített modellre pedig a *11. tábla* tartalmazza.

A két modell különbsége a $-2\ln L$ statisztika tekintetében $27.153-24.781 = 2.372$, mely $df = 3$ szabadsági fok mellett nem szignifikáns (a két modell 3 paraméterben különbözik). A három alternatívaközi hatással történő bővítés

tehát nem javítja jelentősen a likelihood kritériumot, így az egyéb alternatíváktól való függetlenség hipotézise jelen minta esetén fenntartható.

Felhívjuk a figyelmet végül, hogy a Cox-regresszió (Breslow-likelihood) alkalmazása a kiválasztónak megengedi, hogy választási halmazából egyidejűleg ne csak egy, hanem több alternatívát is kiválasszon: adott individuum mellett ennek megfelelően jelenik meg többször a *status* változó $S = 1$ értékkel, $t = 1$ „event time" érték mellett.

| I | Mód | $UI$ | $S$ | $t$ | $D_A$ | $D_R$ | $D_V$ | $D_AUI$ | $D_RUI$ | $D_VUI$ | $D_RAI$ | $D_VRI$ | $D_AVI$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A | 10.0 | N | 2 | 1 | 0 | 0 | 10.0 | .0 | .0 | .0 | .0 | 10.5 |
| 1 | R | 4.5 | I | 1 | 0 | 1 | 0 | .0 | 4.5 | .0 | 10.0 | .0 | .0 |
| 1 | V | 10.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 10.5 | .0 | 4.5 | .0 |
| 2 | A | 5.5 | I | 1 | 1 | 0 | 0 | 5.5 | .0 | .0 | .0 | .0 | 7.5 |
| 2 | R | 4.0 | N | 2 | 0 | 1 | 0 | .0 | 4.0 | .0 | 5.5 | .0 | .0 |
| 2 | V | 7.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 7.5 | .0 | 4.0 | .0 |
| 3 | A | 4.5 | N | 2 | 1 | 0 | 0 | 4.5 | .0 | .0 | .0 | .0 | 5.5 |
| 3 | R | 6.0 | N | 2 | 0 | 1 | 0 | .0 | 6.0 | .0 | 4.5 | .0 | .0 |
| 3 | V | 5.5 | I | 1 | 0 | 0 | 1 | .0 | .0 | 5.5 | .0 | 6.0 | .0 |
| 4 | A | 3.5 | N | 2 | 1 | 0 | 0 | 3.5 | .0 | .0 | .0 | .0 | 5.0 |
| 4 | R | 2.0 | N | 2 | 0 | 1 | 0 | .0 | 2.0 | .0 | 3.5 | .0 | .0 |
| 4 | V | 5.0 | I | 1 | 0 | 0 | 1 | .0 | .0 | 5.0 | .0 | 2.0 | .0 |
| 5 | A | 1.5 | I | 1 | 1 | 0 | 0 | 1.5 | .0 | .0 | .0 | .0 | 4.0 |
| 5 | R | 4.5 | N | 2 | 0 | 1 | 0 | .0 | 4.5 | .0 | 1.5 | .0 | .0 |
| 5 | V | 4.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 4.0 | .0 | 4.5 | .0 |
| 6 | A | 10.5 | N | 2 | 1 | 0 | 0 | 10.5 | .0 | .0 | .0 | .0 | 10.5 |
| 6 | R | 3.0 | I | 1 | 0 | 1 | 0 | .0 | 3.0 | .0 | 10.5 | .0 | .0 |
| 6 | V | 10.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 10.5 | .0 | 3.0 | .0 |
| 7 | A | 7.0 | I | 1 | 1 | 0 | 0 | 7.0 | .0 | .0 | .0 | .0 | 9.0 |
| 7 | R | 3.0 | N | 2 | 0 | 1 | 0 | .0 | 3.0 | .0 | 7.0 | .0 | .0 |
| 7 | V | 9.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 9.0 | .0 | 3.0 | .0 |
| 8 | A | 9.0 | N | 2 | 1 | 0 | 0 | 9.0 | .0 | .0 | .0 | .0 | 9.0 |
| 8 | R | 3.5 | I | 1 | 0 | 1 | 0 | .0 | 3.5 | .0 | 9.0 | .0 | .0 |
| 8 | V | 9.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 9.0 | .0 | 3.5 | .0 |
| 9 | A | 4.0 | I | 1 | 1 | 0 | 0 | 4.0 | .0 | .0 | .0 | .0 | 5.5 |
| 9 | R | 5.0 | N | 2 | 0 | 1 | 0 | .0 | 5.0 | .0 | 4.0 | .0 | .0 |
| 9 | V | 5.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 5.5 | .0 | 5.0 | .0 |
| 10 | A | 22.0 | N | 2 | 1 | 0 | 0 | 22.0 | .0 | .0 | .0 | .0 | 22.5 |
| 10 | R | 4.5 | I | 1 | 0 | 1 | 0 | .0 | 4.5 | .0 | 22.0 | .0 | .0 |
| 10 | V | 22.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 22.5 | .0 | 4.5 | .0 |
| 11 | A | 7.5 | N | 2 | 1 | 0 | 0 | 7.5 | .0 | .0 | .0 | .0 | 10.0 |
| 11 | R | 5.5 | I | 1 | 0 | 1 | 0 | .0 | 5.5 | .0 | 7.5 | .0 | .0 |
| 11 | V | 10.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 10.0 | .0 | 5.5 | .0 |
| 12 | A | 11.5 | N | 2 | 1 | 0 | 0 | 11.5 | .0 | .0 | .0 | .0 | 11.5 |
| 12 | R | 3.5 | N | 2 | 0 | 1 | 0 | .0 | 3.5 | .0 | 11.5 | .0 | .0 |
| 12 | V | 11.5 | I | 1 | 0 | 0 | 1 | .0 | .0 | 11.5 | .0 | 3.5 | .0 |
| 13 | A | 3.5 | I | 1 | 1 | 0 | 0 | 3.5 | .0 | .0 | .0 | .0 | 4.5 |
| 13 | R | 4.5 | N | 2 | 0 | 1 | 0 | .0 | 4.5 | .0 | 3.5 | .0 | .0 |
| 13 | V | 4.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 4.5 | .0 | 4.5 | .0 |
| 14 | A | 12.0 | N | 2 | 1 | 0 | 0 | 12.0 | .0 | .0 | .0 | .0 | 11.0 |
| 14 | R | 3.0 | I | 1 | 0 | 1 | 0 | .0 | 3.0 | .0 | 12.0 | .0 | .0 |
| 14 | V | 11.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 11.0 | .0 | 3.0 | .0 |

*9. táblázat.* Irreleváns alternatívák függetlenségvizsgálata

| I | Mód | $UI$ | $S$ | $t$ | $D_A$ | $D_R$ | $D_V$ | $D_AUI$ | $D_RUI$ | $D_VUI$ | $D_RAI$ | $D_VRI$ | $D_AVI$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | A | 18.0 | N | 2 | 1 | 0 | 0 | 18.0 | .0 | .0 | .0 | .0 | 20.0 |
| 15 | R | 5.5 | I | 1 | 0 | 1 | 0 | .0 | 5.5 | .0 | 18.0 | .0 | .0 |
| 15 | V | 20.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 20.0 | .0 | 5.5 | .0 |
| 16 | A | 23.0 | N | 2 | 1 | 0 | 0 | 23.0 | .0 | .0 | .0 | .0 | 21.5 |
| 16 | R | 5.5 | I | 1 | 0 | 1 | 0 | .0 | 5.5 | .0 | 23.0 | .0 | .0 |
| 16 | V | 21.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 21.5 | .0 | 5.5 | .0 |
| 17 | A | 4.0 | N | 2 | 1 | 0 | 0 | 4.0 | .0 | .0 | .0 | .0 | 4.5 |
| 17 | R | 3.0 | I | 1 | 0 | 1 | 0 | .0 | 3.0 | .0 | 4.0 | .0 | .0 |
| 17 | V | 4.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 4.5 | .0 | 3.0 | .0 |
| 18 | A | 5.0 | N | 2 | 1 | 0 | 0 | 5.0 | .0 | .0 | .0 | .0 | 7.0 |
| 18 | R | 2.5 | N | 2 | 0 | 1 | 0 | .0 | 2.5 | .0 | 5.0 | .0 | .0 |
| 18 | V | 7.0 | I | 1 | 0 | 0 | 1 | .0 | .0 | 7.0 | .0 | 2.5 | .0 |
| 19 | A | 3.5 | I | 1 | 1 | 0 | 0 | 3.5 | .0 | .0 | .0 | .0 | 7.0 |
| 19 | R | 2.0 | N | 2 | 0 | 1 | 0 | .0 | 2.0 | .0 | 3.5 | .0 | .0 |
| 19 | V | 7.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 7.0 | .0 | 2.0 | .0 |
| 20 | A | 12.5 | N | 2 | 1 | 0 | 0 | 12.5 | .0 | .0 | .0 | .0 | 15.5 |
| 20 | R | 3.5 | I | 1 | 0 | 1 | 0 | .0 | 3.5 | .0 | 12.5 | .0 | .0 |
| 20 | V | 15.5 | N | 2 | 0 | 0 | 1 | .0 | .0 | 15.5 | .0 | 3.5 | .0 |
| 21 | A | 1.5 | I | 1 | 1 | 0 | 0 | 1.5 | .0 | .0 | .0 | .0 | 2.0 |
| 21 | R | 4.0 | N | 2 | 0 | 1 | 0 | .0 | 4.0 | .0 | 1.5 | .0 | .0 |
| 21 | V | 2.0 | N | 2 | 0 | 0 | 1 | .0 | .0 | 2.0 | .0 | 4.0 | .0 |

*9. táblázat.* Irreleváns alternatívák függetlenségvizsgálata (folyt.)

| Változó | $B$ | $SE$ | Wald | $df$ | $p$-value | $\exp(B)$ |
|---|---|---|---|---|---|---|
| $D_A$ | -0.738 | 3.059 | 0.058 | 1 | 0.809 | 0.478 |
| $D_R$ | -3.624 | 3.480 | 1.084 | 1 | 0.298 | 0.027 |
| $D_A * UI$ | -2.234 | 1.899 | 1.384 | 1 | 0.239 | 0.107 |
| $D_R * UI$ | -0.101 | 0.686 | 0.022 | 1 | 0.883 | 0.904 |
| $D_V * UI$ | 0.098 | 0.701 | 0.019 | 1 | 0.889 | 1.103 |
| $D_R * AI$ | 0.445 | 0.686 | 0.421 | 1 | 0.517 | 1.560 |
| $D_V * RI$ | -0.532 | 0.635 | 0.703 | 1 | 0.402 | 0.587 |
| $D_A * VI$ | 1.663 | 1.512 | 1.210 | 1 | 0.271 | 5.275 |

-2 Log Likelihood=24.781

*10. táblázat.* Az alternatívák kereszthatásai

| Változó | $B$ | $SE$ | Wald | $df$ | $p$-value | $\exp(B)$ |
|---|---|---|---|---|---|---|
| $D_A$ | 1.716 | 1.805 | 0.904 | 1 | 0.342 | 5.561 |
| $D_R$ | -3.601 | 3.306 | 1.187 | 1 | 0.276 | 0.027 |
| $D_A * UI$ | -0.795 | 0.363 | 4.795 | 1 | 0.029 | 0.451 |
| $D_R * UI$ | 0.122 | 0.590 | 0.043 | 1 | 0.837 | 1.129 |
| $D_V * UI$ | -0.422 | 0.257 | 2.687 | 1 | 0.101 | 0.656 |

-2 Log Likelihood=27.153

*11. táblázat.* Redukált modell alternatíva-közi interakciók nélkül

# 8 Függelék

A túlélési ,,survival" függvény ($t$, $T$: idő)

$$S(t) = \Pr(T \geq t) = 1 - F(t)\,,$$

a ,,hazard rate" függvény:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \ ,$$

Cox-,,proportional hazards":

$$h(t \mid x) = \frac{f(t \mid x)}{S(t \mid x)} = h_0(t) e^{x^T \beta} \ ,$$

ahol $h_0(t)$ a ,,base-line hazard", és a rétegzett ,,hazard" függvény:

$$h_g(t \mid x) = h_{0g}(t) e^{x^T \beta} \ .$$

# 9 Összefoglalás

A tanulmány a diszkrét kiválasztási modell paraméterbecslésének a mikéntjét mutatja be illusztratív példán attól függően, hogy csak az individuum, csak az alternatívák, vagy mindkettő jegyeire egyidejűleg áll rendelkezésre információ. Az individuális változó megfelelő értékét valamennyi döntéshez hozzárendelve, e globalizált változó paramétere közös modellben becsülhető az alternatíva-specifikus változó egyébként is globális paraméterével. Az eljárás feltételes likelihood maximálását igényli, mely a standard Cox-regresszióra visszavezethető. A tanulmány ennek előnyeire irányítja a figyelmet. A Cox-regresszió alkalmazása révén mintavételi következtetésekre, modellszelekcióra is lehetőség nyílik, továbbá vizsgálható az IIA (independence from irrelevant alternatives) hipotézis, az alternatívákhoz kötődő interakciók paramétereinek szeparált, vagy csoportos tesztelésével.

## Irodalom


1. Agresti, A. (2002) *Categorical Data Analysis*. Second Edition, Wiley, New York
2. Greene, W. H. (2003) *Econometric Analysis*. Prantice Hall
3. Guadagni, P. M., Little, J. D. C. (1983) A Logit Model of Brand Choice Calibrated on Scanner Data, *Marketing Science*, Volume 2, Issue 3, 203–238
4. Hajdu, O. (2003) *Többváltozós statisztikai számítások*, KSH, Budapest
5. Hajdu, O. (2004) A csődesemény logit regressziójának kismintás problémái, *Statisztikai Szemle*, 82. évf. 4. sz. 390–422
6. Kleinbaum, D. G., Klein, M. (2002) *Logistic Regression*. Second Edition, Springer, New York
7. Kuhfeld, W. F. (2003) *Marketing Research Methods in SAS*. SAS Institue Inc., NC, USA
8. Norusis, M. J. (2006) *SPSS 13.0*. Advanced Statistical Procedures Companion, Prentice Hall

9. The MDC procedure: http://support.sas.com/rnd/app/papers/mdc.pdf
10. The QLIM procedure: http://support.sas.com/rnd/app/papers/qlim.pdf

## PARAMETER ESTIMATION FOR MULTINOMIAL DISCRETE CHOICE MODEL USING COX-REGRESSION

Considering the so-called ,,multinomial discrete choice" model the focus of this paper is on the estimation problem of the parameters. Especially, the basic question arises how to carry out the point and interval estimation of the parameters when the model is mixed i.e. includes both individual and choice-specific explanatory variables while a standard MDC computer program is not available for use. The basic idea behind the solution is the use of the Cox-proportional hazards method of survival analysis which is available in any standard statistical package and provided a data structure satisfying certain special requirements it yields the MDC solutions desired. The paper describes the features of the data set to be analysed.