MÓDSZERTANI TANULMÁNYOK

A CSŐDESEMÉNY LOGIT-REGRESSZIÓJÁNAK KISMINTÁS PROBLÉMÁI

DR. HAJDU OTTÓ

A tanulmány módszertani útmutatás arra a kismintás esetre, amikor bináris kimenetű változó értékének a bekövetkezési valószínűségét *alacsony* elemszámú minta alapján vagyunk kénytelenek modellezni, adott magyarázóváltozók értékeinek ismeretében. Ekkor ugyanis a konvencionális nagymintás (aszimptotikusan kedvező tulajdonságú) maximum likelihood módszer nem mindig definiálható, de ha definiálható, akkor is félrevezető, torzított eredményt produkálhat. A mintából való statisztikai következtetés speciális módszertani részleteit a dichotom logisztikus regresszió kapcsán mutatom be, de a polichotom esetre is kiterjeszthetők.

TÁRGYSZÓ: Logisztikus regresszió. Feltételes maximum likelihood. Permutációs eloszlás.

A diszkrét kategóriaskálán mért *Y* változó kimenetének az előrejelzését klasszifikációnak nevezzük. Ennek során magyarázóváltozók szintjeinek ismert **x** kombinációja – kovariánsa – mellett kalkuláljuk *Y* kategóriáinak a feltételes valószínűségeit, és a vizsgált *i* megfigyelési egységet a legvalószínűbb kategóriához rendeljük. Például egy hitelkérelem minősítése során, csődkockázati szempontból kockázatosként vagy kockázatmentesként minősíthetünk egy gazdasági egységet (többek között mérlege és eredménye, tevékenységi köre, működési formája, stb. ismeretében) a döntés pénzügyi következményeivel együtt. A logisztikus regresszió a klasszifikálás egyik klasszikus módszere, így alkalmazása a csődkockázat mérésében is kézenfekvő.

Ha az eredmény jellegű (dependent, response) változó bináris, vagyis két lehetséges kimenete „1" és „0", „igen/nem", akkor dichotom (binomiális) logisztikus regresszióról beszélünk. A függő változó eloszlásának az ismeretében a logisztikus regresszió paramétereinek a becslésére a maximum likelihood (ML) módszer kínálkozik, viszont a maximum likelihood eljárás kedvező tulajdonságai (például minimum variancia, konzisztencia) aszimptotikusan, *nagymintás* esetben érvényesülnek. Ugyanakkor a csődhelyzet klasszifikálása a kismintás következtetés tipikus esete, hiszen a csődesemény *relatíve* ritka jelenség. Kiváltképp alacsony gyakoriságú bizonyos tevékenységi körökben, (szak)ágazatokban, tehát egy szakágazati szintre lebontott „csődmodell" kismintás becslése kényszerű adottság. Jelen tanulmány alapvető célja, hogy a csődkockázat mérése kapcsán a logisztikus regresszió ML becslési problémáira fölhívja a figyelmet, és fölismerésükre, kezelésükre megfelelő módszertant javasoljon.

A feltétel nélküli maximum likelihood eljárás alkalmazása szempontjából alapvető probléma a *kiegyensúlyozatlan* minta esete, melyben (tekintet nélkül a mintanagyságra) relatíve nagyon alacsony (akár 5 százalék alatti) a csődesemények aránya, másfelől a *szeparált* minta esete, melyben a csődesemény egyértelműen a magyarázó változó egy adott szegmenséhez, a komplementer „működő" események pedig egy jól elhatárolt, másik szegmenséhez tartoznak. Míg az előbbi esetben van egyedi ML-megoldás, de az torzított és magas mintavételi varianciával bír, addig az utóbbi esetben *nem is létezik* a ML-megoldás. A harmadik lényeges problémát az okozza, mikor *a priori* információnk van a csődesemények arányáról a sokaságban (ez az információ a nemzetgazdaságban rendelkezésre áll) és ez az arány jelentősen eltér a megfelelő mintabeni aránytól, további torzítást okozva a paraméterek becslésében.

A ritka „1" esemény kezelését az aszimptotikus logisztikus regresszió megfelelő korrekcióval való alkalmazása, vagy a csőd/működés események egzakt permutációin alapuló ún. egzakt (nem aszimptotikus) logisztikus regresszió (ELR) egyaránt szolgálja. Az ELR-eljárás a regressziós paraméterek *elégséges statisztikáinak* az egzakt, feltételes, permutációs eloszlásán alapuló módszertana. Mikor az aszimptotikus ML-becslés nem létezik, az ELR-módszer használatával akkor is következtetni tudunk a regressziós paraméterekre.

Jelen tanulmány az eredmények értelmezése végett előbb áttekinti magát a döntési problémát, amely döntés érdekében a döntéshozó regressziós megalapozásra támaszkodik. Ezt követően foglalkozunk a kiegyensúlyozatlanság, torzítottság és a szeparáltág problémáival, majd az egzakt logisztikus regresszió módszertanának elméleti részleteit tárgyaljuk dichotom Y esetén. Ennek során olyan gyakorlati példákon követjük nyomon a statisztikai következtetés (hipotézisek tesztelése, pont- és intervallumbecslés) menetét, sajátosságait, melyek az aszimptotikus ML-módszerrel nem elemezhetők. Végül néhány, a becslések torzítottságát kezelő algoritmust ajánlunk az elemzők figyelmébe. Az illusztratív példák csődbement és működő gazdasági vállalkozások klasszifikálását tárgyalják, a mindenkori módszertani mondanivalóhoz igazodó adatállományok alapján.[1]

1. A DICHOTOM DÖNTÉSI MODELL

Tekintsük független, bináris $Y_i=\{1,0\}$ változók (i=1,2,...,n) sorozatát, amely változók kimenete az $x_1, x_2,..., x_p$ magyarázó változók szintjeinek valamely $\mathbf{x}_k=(x_{k1}, x_{k2},..., x_{kp})$ rögzített kombinációja mellett következik be. Az $\mathbf{x}_k$ kombinációt *kovariánsnak* nevezzük, és adott kovariáns mellett több megfigyelést is végezhetünk. Az Y=1 kimenet adott tulajdonság meglétét, Y=0 pedig a hiányát jelzi. Esetünkben Y=1 gazdasági vállalkozások „*csődjét*", Y=0 pedig „*működését*" jelenti. Jelölje π_x a Pr(Y=1|**x**) esemény feltételes valószínűségét, mely a $\pi_x/(1-\pi_x)$ ún. odds-arány alapján

$$\pi_x = \frac{\pi_x/(1-\pi_x)}{1+\pi_x/(1-\pi_x)} = \frac{odds_x}{1+odds_x}. \qquad /1/$$

[1] A számítások a SAS-programmal készültek.

A logisztikus regresszió szerint az odds-arány logaritmusa (egyben a π_x valószínűség *logitja*) az

$$\ln\left(odds_x\right) = \operatorname{logit}\left(\pi_x\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad /2/$$

lineáris „prediktor” szerint alakul, mellyel

$$\pi_x = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}} = \frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{1+e^{\boldsymbol{\beta}'\mathbf{x}}} = \frac{1}{1+e^{-\boldsymbol{\beta}'\mathbf{x}}}, \qquad /3/$$

ahol

$$\boldsymbol{\beta}' = \left(\beta_0, \beta_1, \ldots, \beta_p\right)$$

az ismeretlen regressziós koefficiensek vektora és x_0 a tengelymetszethez rendelt összegző vektor. Értelemszerűen a komplementer esemény valószínűsége

$$1 - \pi_x = \frac{1}{1+e^{\boldsymbol{\beta}'\mathbf{x}}} = \frac{e^{-\boldsymbol{\beta}'\mathbf{x}}}{1+e^{-\boldsymbol{\beta}'\mathbf{x}}}.$$

A regressziós paraméterek értelmezését az e^{β_j} inflátor (deflátor) faktor szolgálja, mely az x_j magyarázó változó egységnyi *abszolút* növekményének az odds-arányra gyakorolt *multiplikatív* hatását mutatja, a többi magyarázó változó szinten tartása mellett:

$$odds_{x_j+1} = odds_{x_j} \cdot e^{\beta_j}. \qquad /4/$$

Ha a $\Delta x_j=1$ változásnak a csődvalószínűségre gyakorolt hatását mérjük, akkor előbb felírható a $\boldsymbol{\beta}'\mathbf{x}$ szerinti derivált

$$\frac{\partial \pi_x}{\partial\left(\boldsymbol{\beta}'\mathbf{x}\right)} = \frac{\partial \frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{1+e^{\boldsymbol{\beta}'\mathbf{x}}}}{\partial\left(\boldsymbol{\beta}'\mathbf{x}\right)} = \pi_x\left(1-\pi_x\right), \qquad /5/$$

ahonnan

$$\frac{\partial \pi_x}{\partial x_j} = \beta_j \pi_x\left(1-\pi_x\right).$$

Az előrejelzés érdekében a $\boldsymbol{\beta}'$ regressziós paramétereket egy $y_1, y_2, \ldots, y_n$ független, véletlen minta alapján becsülnünk kell, majd a becslések birtokában Y előrejelzése egy döntési kritérium alapján történik, az alábbiak szerint. A π_x valószínűség magas vagy alacsony voltának az elhatárolásához rögzítünk egy alkalmasan megválasztott kritikus C_π

„cut-off-value” értéket, és e kritikus érték alapján az előrejelzés: $\hat{Y} = 1 |\ \pi_x \geq C_\pi$, egyébként az előrejelzés 0. A C_π érték rögzítését a mintán verifikált „klasszifikációs” mátrix és a veszteség (haszon) függvény együtt segíti:

$$Loss = c_{11}D_{11} + c_{10}D_{10} + c_{01}D_{01} + c_{00}D_{00},$$

ahol D_{11} és D_{00} a korrekt, D_{10} és D_{01} az inkorrekt előrejelzések gyakorisága, $c_{y\hat{y}}$ pedig a döntéssel járó fajlagos költség (haszon) koefficiens. Pozitív c_{10} és c_{01} választással veszteséget minimálunk, míg pozitív c_{11} és c_{00} választással hasznot maximálunk. Speciálisan, ha $c_{11} = c_{00} = 0$, és $c_{10} = c_{01} = 1$, akkor az összes inkorrekt klasszifikáció gyakoriságát minimalizáljuk.

Döntési szabályként azt a kritikus értéket célszerű választani, amely mellett a veszteség minimális, vagy a haszon maximális. Mindazonáltal célszerű figyelembe venni, hogy általában (különösen a csődesély minősítésekor) a kétféle inkorrekt előrejelzés nem egyforma pénzügyi következményű, és adott kimenet mellett (például csődbement a vállalkozás) a korrekt és az inkorrekt klasszifikáció pénzügyi következményei nem föltétlenül zérus összegűek. Ezt az aszimmetriát illusztrálja egy hitelt nyújtó szempontjából az 1. tábla költségmátrixa egységnyi hitel odaítélése felől történő döntés során, miközben veszteséget kíván minimalizálni.

1. tábla

Egyféle változat egységnyi hitel nyújtásának pénzügyi veszteségeiről

Tény	Előrejelzés	
	Csőd	Működés
Csőd	0	1
Működés	0	–0,2

A tábla azt sugallja, hogy a döntést elősegítő kritikus „cut-off-value” értéket nem gyakorisági, hanem pénzügyi alapon indokoltabb behatárolni. A kölcsön nyújtását szigoríthatjuk (lazíthatjuk), ha a „csőd/működés” hibás döntés egységnyi veszteségét felnagyítjuk (kicsinyítjük) a „működés/csőd” hiba zéró veszteségéhez képest.

Hangsúlyozzuk, hogy a „cut-off-value” értékének a rögzítésével tulajdonképpen a modell illeszkedésének a jóságát befolyásoljuk „korrekt-klasszifikálás” értelemben, aminek a javítása konkrét minta esetén igényelhet olyan magyarázó változót, melyet hipotézisvizsgálat alapján egyébként kizárnánk a modellből.

2. A LOGISZTIKUS REGRESSZIÓ KISMINTÁS KÉRDÉSEI

Kismintás esetben a logisztikus regresszió alkalmazása számos becslési és hipotézisvizsgálati problémát vet fel. A kismintás probléma mind a teljes mintanagyság, mind az „1” egyedek relatív számossága tekintetében értelmezhető.

2.1. A ritka esemény problémája

A csődvalószínűség modellezésének alternatív, de a fentivel ekvivalens megközelítését teszi lehetővé a *logisztikus eloszlás* alapján való döntés, az alábbiak szerint. Tekintsük az Y^* folytonos, de közvetlenül nem megfigyelhető (latens) „csődmérték"-változót, amelynek *x* feltétel mellett várható értéke η_x. A logisztikus eloszlás sűrűségfüggvénye ekkor:

$$\text{Logistic}\left(Y^* \mid \eta_x\right) = \frac{e^{-\left(Y^* - \eta_x\right)}}{\left(1 + e^{-\left(Y^* - \eta_x\right)}\right)^2},$$

ahol

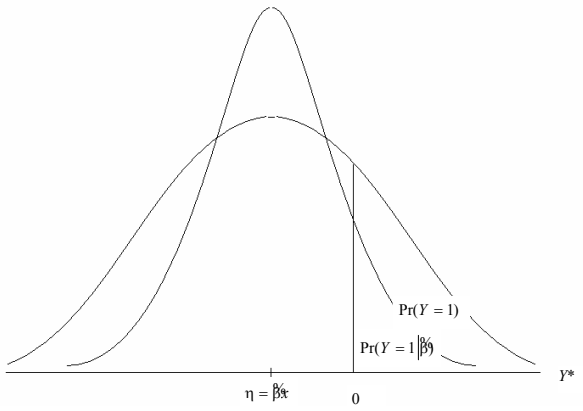
$$\eta_x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p .$$

Képezve a

$$\Pr\left(Y^* > 0\right) = \int_0^\infty \text{Logistic}\left(Y^* \mid \eta_x\right) dY^* = \frac{1}{1 + e^{-\eta_x}} = \frac{e^{\eta_x}}{1 + e^{\eta_x}} \qquad /6/$$

kumulatív valószínűséget, ha az Y^* változót úgy diszkretizáljuk, hogy az $Y^*>0$ eseményt „1", a komplementer eseményt pedig „0" definiálja, akkor /3/ és /6/ láthatóan ekvivalens valószínűségi modellek. Ez arra hívja fel a figyelmet, hogy *kiegyensúlyozatlan* minta esetén, mikoris az „1" egyedek ritkán fordulnak elő a mintában (*f*/*n* relatív gyakoriságuk kicsiny, akár kisebb mint 5 százalék), akkor a π_x valószínűségnek egy $\tilde{\beta}$ pontbecslésből származó $\tilde{\pi}_x$ pontbecslése akkor is alulbecsült, ha $\tilde{\beta}$ egyébként torzítatlan becslése a megfelelő regressziós paraméternek (*King–Zeng* [2001a]). Ezt illusztrálja az 1. ábra.

1. ábra. A feltételes valószínűség alulbecslése



Az ábrán a nagyobb szórású logisztikus sűrűségfüggvény a populáció eloszlását írja le a latens „csődmérték" változó tekintetében. Így ezen görbe alatt az $Y^*=0$ értéktől jobbra lévő terület a Pr(Y=1) sokasági valószínűséget jelenti. A sokasági szórást az egyelemű minta standard hibája reprezentálja. A többelemű mintavétel eredményeként nyert torzítatlan $\tilde{\beta}$ becslések által generált eloszlás szükségszerűen alacsonyabb szórású, és ezt az ábrán a csúcsosabb függvény írja le. Az alacsonyabb szórású esetben láthatóan kisebb az $Y^*=0$ értéktől jobbra eső terület, vagyis a $\Pr(Y=1|\tilde{\beta})$ valószínűség. A $\tilde{\pi}_x$ pontbecslés tehát az „1" esemény valószínűségét alulbecsli.

2.2. Aszimptotikus, torzított paraméterbecslés

A regressziós paraméterek becslése és tesztelése mind a legkisebb négyzetek elvén, mind a maximum likelihood módszeren alapulhat. Tekintsünk egy n-elemű y_i (i=1,2,...,n) független mintát, melyben n_k számú megfigyelés tartozik az x_k kovariánshoz, és ezek között f_k az „1" tulajdonságúak gyakorisága.

Az *iterative újrasúlyozott legkisebb négyzetek* módszere a

$$\sum_k \frac{1}{n_k \hat{\pi}_k \left(1-\hat{\pi}_k\right)} \left(f_k - n_k \hat{\pi}_k\right)^2 \to \min \qquad /7/$$

súlyozott négyzetösszeget minimálja, ahol adott becslés birtokában a súly újraszámításra kerül új paraméterekhez vezetve mindaddig, míg az eredmények nem változnak jelentősen.

Természetesen a *maximum likelihood* elv alkalmazása is kézenfekvő, hiszen egzakt ismeretünk van az eredményváltozó eloszlását illetően, mely *Bernoulli*-folyamatot követ. Pontbecsléskor a minta együttes likelihoodját maximáljuk, melyet súlyozatlan formában az alábbi szorzat definiál

$$L = \Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = \prod_{\{y_i=1\}} \pi_i \prod_{\{y_i=0\}} \left(1-\pi_i\right) = \prod_{i=1}^{n} \pi_i^{y_i} \left(1-\pi_i\right)^{1-y_i} .$$

A /2/ logit modellt behelyettesítve, a likelihood értékét többféle formában is felírhatjuk attól függően, hogy melyik formula használata egyszerűsíti inkább a számításokat:

$$L = \prod_{i=1}^{n} \frac{\left(odds_i\right)^{y_i}}{1+odds_i} = \frac{e^{\sum_{i=1}^{n} y_i \sum_{j=0}^{p} \beta_j x_{ij}}}{\prod_{i=1}^{n} \left(1+e^{\sum_{j=0}^{p} \beta_j x_{ij}}\right)} = \frac{e^{\sum_{j=0}^{p} \beta_j \sum_{i=1}^{n} y_i x_{ij}}}{\prod_{i=1}^{n} \left(1+e^{\sum_{j=0}^{p} \beta_j x_{ij}}\right)} = \frac{e^{\boldsymbol{\beta}'\mathbf{t}}}{\prod_{i=1}^{n} \left(1+e^{\boldsymbol{\beta}'\mathbf{x_i}}\right)}, \qquad /8/$$

ahol a $\mathbf{t} = \left(t_0, t_1, ..., t_p\right)'$ vektor általános eleme

$$t_j = \sum_{i=1}^{n} y_i x_{ij} \quad \left(j = 0, 1, 2, ..., p\right)$$

egyben a β_j paraméter ún. *elégséges statisztikája*, mely jelen tanulmány központi fogalma.[2] Mivel *y* értéke csak 1 vagy 0 lehet, ezért a t_j statisztika értéke az x_j magyarázó változó y=1 esetekben felvett mintabeli értékeinek az *összege*. Például β_0 elégséges statisztikája t_0, mely az „1" esemény f előfordulási gyakorisága a mintában:

$$t_0 = f\ .$$

Így a „log-likelihood"

$$\ln L = \sum_{i=1}^{n} y_i \ln\left(odds_i\right) + \sum_{i=1}^{n} \ln \frac{1}{1+odds_i} = \sum_{j=0}^{p} \beta_j \left(\sum_{i=1}^{n} y_i x_{ij} \right) + \sum_{i=1}^{n} \ln\left(1-\pi_i\right) =$$

$$= \sum_{j=0}^{p} \beta_j t_j + \sum_{i=1}^{n} \ln\left(1-\pi_i\right). \qquad /9/$$

Ekkor a /9/ kifejezés alapján képzett $\partial \ln L / \partial \beta_j = 0$ maximum-likelihood egyenletrendszer – felhasználva közben a /5/ azonosságból származó $\partial \pi_x / \partial \beta_j = x_j \pi_x \left(1-\pi_x\right)$ deriváltat is – a

$$t_j = \sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \pi_i x_{ij} \quad \left(j = 0,1,2,...,p\right) \qquad /10/$$

módon írható fel.

A /8/ maximálási feladat numerikus megoldása egyben a /7/ minimálási feladatét is megadja (lásd *Jennrich–Moore* [1975]). A *Fisher-scoring* módszert alkalmazva, a becsült paraméterekben történő ellépésvektort az alábbi formula határozza meg:

$$\Delta\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{z}\ ,$$

ahol a $\mathbf{z}$ vektor általános eleme $\left(f_k - n_k\hat{\pi}_k\right) / n_k\hat{\pi}_k\left(1-\hat{\pi}_k\right)$, a súlymátrix általános eleme pedig

$$W_{kk} = n_k \pi_k \left(1-\pi_k\right).$$

A becsült paraméterek aszimptotikus variancia-kovariancia mátrixa ekkor (az általános lineáris modell paraméterbecslésének megfelelően) a Fisher-féle információs mátrix inverze, amely most (*Garthwaite–Jolliffe–Jones* [1995] 245. old.):

$$\mathbf{C}_{\hat{\beta}} = \left[\sum_{i=1}^{n} \pi_i \left(1-\pi_i\right) \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = \left[\sum_{k} n_k \pi_k \left(1-\pi_k\right) \mathbf{x}_k \mathbf{x}_k' \right]^{-1}, \qquad /11/$$

[2] Nem tévesztendő össze a klasszikus Student-*t* statisztikával. Az elégséges statisztika fogalmát lásd *Hunyadi* [2001] vagy *Garthwaite–Jolliffe–Jones* [1995]. Hozzátesszük, hogy a későbbiek megértése nem igényli az elégséges statisztika pontos definiálását.

illetve mátrixformában

$$\mathbf{C}_{\hat{\beta}} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}.$$

Alapvető probléma, hogy /8/ maximálása a paraméterek *torzított* becslését eredményezi bármilyen véges mintában, akkor is, ha egyébként a minta kiegyensúlyozott. A torzítás mértéke a mintanagyság növelésével csökken, és az irodalom szerint n=200 fölött elhanyagolhatóvá válik (*Schaefer* [1983]). *McCullagh* és *Nelder* [1989] megmutatták, hogy a torzítás mértéke bármely általános lineáris modellre az alábbiak szerint számítható:

$$Bias\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{x}, \qquad /12/$$

ahol $\xi_k = -0,5\mu_k'' Q_{kk} / \mu_k'$, és az általános lineáris modell klasszikus jelöléseinek megfelelően az eredményváltozó várható értéke $\mu_k = E\left(Y_k\right)$, a lineáris prediktor $\eta_k = \boldsymbol{\beta}'\mathbf{x}_k$, továbbá μ_k' és μ_k'' az első és másodrendű deriváltak η_k tekintetében, végül Q_{kk} az $\mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'$ ún. *hat* mátrix megfelelő diagonális eleme.

Mindemellett, mivel a csődhelyzet elemzése során nem a paramétereken, hanem a belőlük számított odds-arányon és csődvalószínűségen van a hangsúly, ezek a jellemzők (a nemlineáris) átvitel miatt akkor is torzítottak lennének, ha maguk a paraméterek egyébként torzítatlanok.

A ML-becslés alkalmazása szempontjából még kritikusabb probléma, hogy bizonyos esetekben véges, egyedi ML-megoldás *nem is létezik*.

2.3. Szeparáltság és átfedés

Egyedi, véges maximum likelihood becslés nem létezik akkor, ha a megfigyelések a magyarázó változók bármelyike tekintetében *teljesen*, vagy *kvázi* módon *szeparáltak* (*Albert–Anderson* [1984]). A problémát az alábbi példa világítja meg.

Egyetlen magyarázó változó esetén, ha valamennyi csődbe ment vállalkozás veszteséges (negatív az eredménye) és valamennyi működő vállalkozás nyereséges (pozitív az eredménye), akkor a vállalkozások teljesen szeparáltak. A zéró nyereség mint szeparáló érték minden vállalkozást korrekten klasszifikál. Ha eközben zéró eredményt mind a csődbe ment, mind a működő vállalkozások között megengedünk, akkor a vállalkozások, úgymond, kváziszeparáltak. Két magyarázóváltozót tekintve, ha a vállalkozásokat az eredményük és a likviditásuk tekintetében a síkban ábrázoljuk, és húzható egy olyan egyenes, melynek egyik oldalán csak csődbe ment, másik oldalán pedig csak működő vállalkozások vannak, akkor a vállalkozások teljesen szeparáltak.

Általánosságban az $y_1, y_2, \ldots, y_n$ minta *teljesen szeparált*, ha léteznek a_0, a_1, $a_2, \ldots$, a_p konstansok, melyek közül legalább egy pozitív indexű nem zéró, és

$$a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} > 0$$

minden y_i=0 esetre, és

$$a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} < 0$$

minden y_i=1 esetre. Ugyanakkor az $y_1, y_2, \ldots, y_n$ minta *kváziszeparált*, ha

$$a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} \geq 0$$

minden y_i=0 esetre, és

$$a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} \leq 0$$

minden y_i=1 esetre. Ha a mintában sem teljes, sem kváziszeparáltság nem található, akkor a minta *átfedéses*. E konfigurációk következménye a ML-megoldás létezésére a következő.

– Ha a mintabeli megfigyelések teljesen szeparáltak, akkor nem létezik egyedi véges megoldás a ML normál egyenletekre. Ha a likelihood függvényt maximáló iterációs eljárást mégis folytatjuk, a loglikelihood zéróhoz csökken, a paraméterek szóródási mátrixa pedig nemkorlátossá válik.

– Ha a mintabeli megfigyelések kváziszeparáltak, akkor nem létezik egyedi véges megoldás a ML normál egyenletekre. Ha a likelihood függvényt maximáló iterációs eljárást mégis folytatjuk, akkor a loglikelihood egy nemzéró konstanshoz csökken, a paraméterek szóródási mátrixa pedig nemkorlátossá válik.

– Ha a mintabeli megfigyelések átfedésesek, akkor létezik egyedi véges megoldás a ML normál egyenletekre.

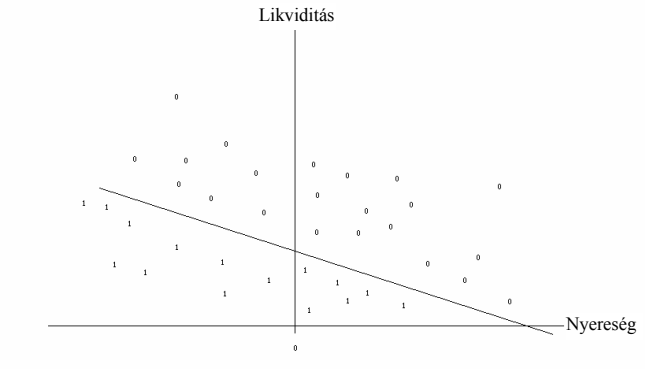
Két magyarázóváltozó esetén a szeparáltság és átfedésesség problémáját illusztrálja a 2. és a 3. ábra. Az ábrák mutatják, hogy hiába vannak átfedések mind x_1, mind x_2 tekintetében külön-külön, a (teljes vagy kvázi-) szeparáltság esetével állunk szemben. Ha bármelyik magyarázóváltozó tekintetében fennáll legalább a kvázi szeparáltság, vagyis az átfedés hiánya, akkor ez elégséges feltétel az egyedi, véges ML-módszer nemlétezéséhez, de hiába van átfedés akár mindegyik magyarázóváltozó tekintetében is külön-külön, ez önmagában nem elégséges feltétel a véges, egyedi ML-megoldás létezéséhez.[3] *Santner* és *Duffy* [1986] ad egy lineáris programozáson alapuló algoritmust azt meghatározandó, hogy a ML-becslés mikor nem létezik.

Fönntartással kell fogadni mindenképpen a ML-elven alapuló következtetéseket akkor is (*King–Ryan* [2002]), ha a véges ML-becslés létezik ugyan, de

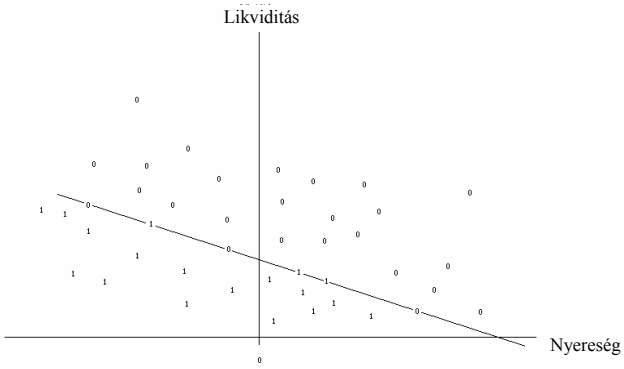
– *ritkán* fordulnak elő vagy az „1”, vagy a „0” egyedek (zéróközeli az arányuk) a mintában,
– csekély mértékű az „1” és „0” egyedek *átfedése* a mintában.

[3] Egyféle empirikus közelítés a szeparáltság meglétének az ellenőrzésére a következő lehet. A log-likelihood maximálása során, ha nyolc iterációs lépésen belül az algoritmus konvergál, akkor nem ellenőrizzük a szeparáltságot. A nyolcadik iterációs lépést követően valamennyi megfigyelésre meghatározzuk az ő megfelelő feltételes valószínűségét. Ha ez minden megfigyelésre 1, akkor az adatok teljesen szeparáltak, a maximálási eljárást megállítjuk. Ha teljes szeparáltság nincs a mintában, de egy megfigyelésre extrém nagy valószínűség (nagyobb vagy egyenlő mint 0,95) adódik, akkor két lehetőség van. Egyfelől lehet átfedéses a minta, és ekkor a maximálási eljárás leáll, ha elérte a maximumot. Másfelől, az adatok lehetnek kváziszeparáltak, ekkor a szóródási mátrix nem korlátos. Ezt a helyzetet jelzi, ha a standardizált magyarázóváltozók szóródási mátrixa valamennyi diagonális eleme meghaladja az 5000 értéket.

2. ábra. Teljesen szeparált megfigyelések két magyarázóváltozó síkjában



3. ábra Kváziszeparált megfigyelések két magyarázóváltozó síkjában



Az ML-egyenletrendszer megoldhatóságának a kérdése az elégséges *t*-statisztika lehetséges terjedelmének az oldaláról is megközelíthető.

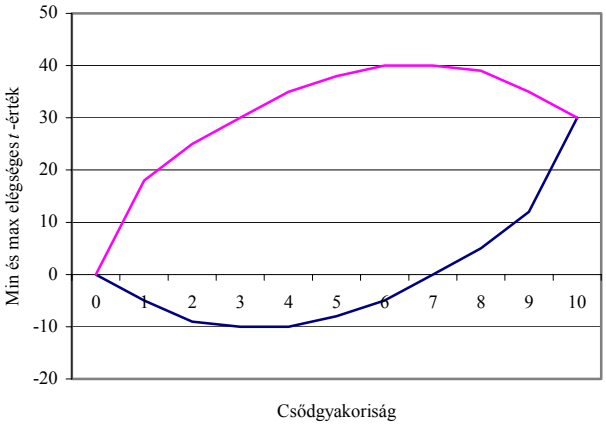
2. tábla

A t_1 *elégséges statisztika határai f számú „1” esemény mellett, tízelemű mintában*

$f=t_0$	x	Kumuláns	
		t_1 alsó határ	t_1 felső határ
0		0	0
1	–5	–5	18
2	–4	–9	25
3	–1	–10	30
4	0	–10	35
5	2	–8	38
6	3	–5	40
7	5	0	40
8	5	5	39
9	7	12	35
10	18	30	30

Tekintsünk 10, a veszteségük tekintetében sorba rendezett gazdálkodó egységet. (Az adatokat a 2. tábla közli.) Ha a 10 elemű mintában például négy „1” tulajdonságú (csőd) cég található, akkor t_1 értéke (támaszkodva x rendezettségére) legalább –10, de legfeljebb 35. Most, ha egy konkrét mintában négy csődbement mellett t_1 értéke éppen –10, vagy éppen 35, akkor véges, egyedi ML-megoldás nem létezik. Ábrázoljuk a 4. ábrán látható módon t_0 függvényében t_1 alsó és felső határát, mely a $0 \le t_0 \le n$ tartományon egy ún. *konvex kiterjesztést* alkot. Akkor van egyedi megoldása a ML-egyenletrendszernek, ha a t_1 statisztika mintabeli értéke e konvex kiterjesztés *belső pontja*.

4. ábra. Konvex kiterjesztés



Világos, hogy az elégséges t-statisztika akkor veszi fel a szélső értékeit, ha a csődbement „1” vállalkozások az **x** szerinti rangsorban mind egymást követve legalul, vagy mind egymást követve legfelül helyezkednek el. Ez pedig a (kvázi- vagy teljes) szeparáltság esete.

A 2. tábla adatait használva, az egyedi ML-becslés nemlétezését illusztrálja az 5. ábra, 4 teljesen szeparált csődeseményt feltételezve az eloszlás felső szegmensén a 10 elemű mintában: **y**=(0,0,0,0,0,0,1,1,1,1). Ekkor az elégséges t_1-statisztika egybeesik a felső határával, azaz $t_1 = 35$. A megoldandó ML-egyenletrendszer /10/ alapján most:

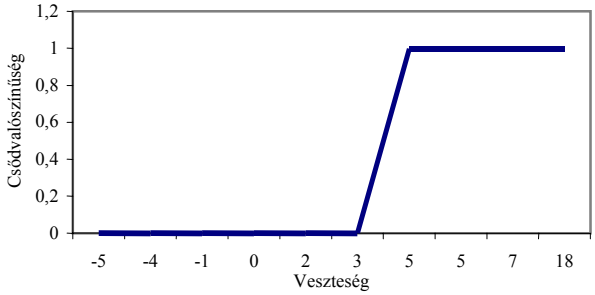
$$t_0 = \sum_{i=1}^{10} y_i = 4 = \sum_{i=1}^{10} \pi_i$$

$$t_1 = \sum_{i=1}^{10} y_i x_i = 35 = \sum_{i=1}^{10} \pi_i x_i \ .$$

A 2. táblát tekintve látható, hogy ez az egyenletrendszer végtelen sok olyan β_0, β_1 paraméterpáros mellett teljesül, melyek a 5. ábrának megfelelően az első 6 megfigyeléshez közel zéró, az utolsó 4 megfigyeléshez pedig közel 1 valószínűséget becsülnek. (Az olva-

só kipróbálhatja például a β_0=–50, β_1=12,6568, vagy a β_0=–55, β_1=13,9225 paraméterekkel.) Ekkor a likelihood 1-hez, a loglikelihood pedig zéróhóz konvergál.

5. ábra. Teljesen szeparált csődesemények becsült valószínűségei



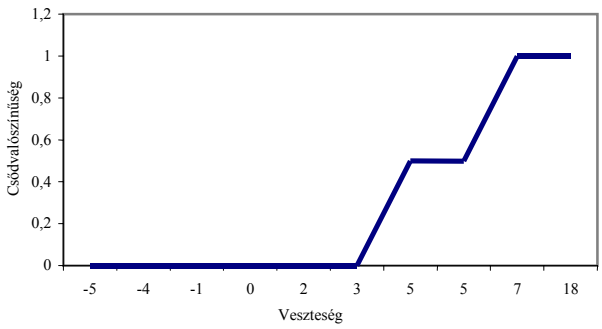
Újra a 2. tábla adatait használva, tekintsünk most egy „háromcsődős", kváziszeparált esetet olymódon, hogy az első hét megfigyelés y=0, az utolsó három viszont y=1 tulajdonságú. Így, mivel x_7=x_8=5, a minta kvázi-szeparált. Egyedi ML-becslés ebben az esetben sem létezik, mert az elégséges t_1-statisztika most is egybeesik a felső határával, ami t_1=30. Mivel t_0=3, ezért a megoldandó ML-egyenletrendszer a következő:

$$t_0 = \sum_{i=1}^{10} y_i = 3 = \sum_{i=1}^{10} \pi_i$$

$$t_1 = \sum_{i=1}^{10} y_i x_i = 30 = \sum_{i=1}^{10} \pi_i x_i \ .$$

A 2. táblát tekintve látható, hogy ez az egyenletrendszer végtelen sok olyan β_0, β_1 paraméterpáros mellett teljesül, melyek az első 6 megfigyeléshez közel zéró, az utolsó 2 megfigyeléshez közel 1 valószínűséget, a 7. és a 8. megfigyelésekhez pedig egyaránt 0,5 közeli valószínűséget becsülnek. (Az Olvasó kipróbálhatja például a (β_0=–38, β_1=7,6), vagy a (β_0=–45, β_1=9) paraméterekkel.) Most a likelihood a $0{,}5^2$, a –2*loglikelihood célfüggvény pedig a 2,773 értékhez konvergál. Az elmondottakat a 6. ábra szemlélteti.

6. ábra. Kváziszeparált csődesemények becsült valószínűségei



Mikor a tengelymetszet tekintetében nem, csak a regressziós meredekség tekintetében kell optimálnunk, akkor természetesen rögzített β_0 mellett már létezik egyedi ML-becslés a β_1 paraméterre, hiszen a csődvalószínűség β_1 tekintetében szigorú monoton változik. Ha például a „négycsődös" teljesen szeparált minta esetén β_0 rögzített értéke –0,40547 (az x=0 melletti ML-becslés), akkor e feltétel mellett β_1 ML becslése 0,413, és a β_1 paraméterhez tartozó ML-egyenlet más becslés mellett nem teljesül. Ugyanebben a példában, ha β_0 rögzített értéke zéró, akkor β_1 ML becslése 0,375. Ebben az értelemben a tengelymetszetet *zavaró*, „*nuisance*" paraméterként is szokás kezelni.

3. EGZAKT LOGISZTIKUS REGRESSZIÓ

Abban az esetben, mikor /8/ definiálható, és a tengelymetszetre való következtetés nem célunk, a becslést alapozhatjuk az aszimptotikus, de *feltételes* maximum likelihood módszerre. Ha /8/ nem definiálható, akkor egyetlen lehetséges megoldás az $\mathbf{y}=(y_1, y_2, ..., y_n)$ mintabeli szekvencia minden lehetséges permutációján alapuló *egzakt* módszert használni.

3.1. Feltételes, egzakt permutációs likelihood

Ha célunk a parciális regressziós paraméterek egy szűk csoportjára való következtetés, akkor a többi paraméter – elégséges statisztikáik rögzítése révén – eliminálható a likelihood függvényből a következők szerint. Tekintsük az elégséges statisztikák $\mathbf{t} = \left(t_0, t_1, t_2, ..., t_p\right)'$ vektorát a mintában, ahol a korábbiaknak megfelelően

$$t_j = \sum_{i=1}^{n} y_i x_{ij} \ . \qquad /13/$$

A minta /8/ likelihoodjának a felhasználásával az elégséges statisztikák együttes eloszlása:

$$\Pr\left(\mathbf{T} = \mathbf{t}\right) = \frac{c(\mathbf{t}) e^{\boldsymbol{\beta}'\mathbf{t}}}{\prod_{i=1}^{n}\left(1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}\right)},$$

ahol $c(\mathbf{t})$ mindazon $\mathbf{y}$ szekvenciák száma (count), melyek éppen a $\mathbf{t}$ vektort eredményezik. Partícionáljuk most a magyarázó változókat az $\mathbf{X}=[\mathbf{X}_0, \mathbf{X}_1]$ módon két csoportba, és legyen feladatunk az $\mathbf{X}_1$ változók $\boldsymbol{\beta}_1'$ paramétereire való következtetés a $\mathbf{t}_1$ elégséges statisztikáik alapján. Ennek érdekében tekintsük a $\sum_{i=1}^{n} y_{iR} x_{ij} = u_j$ jellegű összeget a mintatér egy másik $\mathbf{y}_R$ szekvenciáján is, megfelelő értékeit foglaljuk az $\mathbf{u} = \left(\mathbf{u}_0, \mathbf{u}_1\right)'$ vektorba, majd képezzük a $\mathbf{t}_0$ elégséges statisztikák bekövetkezésének az együttes valószínűségét:

$$\Pr\left(\mathbf{T}_0 = \mathbf{t}_0\right) = \sum_u \frac{c(\mathbf{u}_1, \mathbf{t}_0) e^{\boldsymbol{\beta}_1' \mathbf{u}_1 + \boldsymbol{\beta}_0' \mathbf{t}_0}}{\prod_{i=1}^{n} \left(1 + e^{\boldsymbol{\beta}' \mathbf{x}_i}\right)},$$

ahol $c(\mathbf{u}_1, \mathbf{t}_0)$ mindazon **y** vektorok száma melyekre $\mathbf{X}_1\mathbf{y}=\mathbf{u}_1$ és $\mathbf{X}_0\mathbf{y}=\mathbf{t}_0$. Ekkor az elégséges statisztikák *feltételes* együttes likelihoodja:

$$\Pr\left(\mathbf{T}_1 = \mathbf{t}_1 \mid \mathbf{T}_0 = \mathbf{t}_0\right) = \frac{\Pr\left(\mathbf{T} = \mathbf{t}\right)}{\Pr\left(\mathbf{T}_0 = \mathbf{t}_0\right)} = L\left(\mathbf{t}_1 \mid \boldsymbol{\beta}_1'\right)$$

$$= \frac{c(\mathbf{t}) e^{\boldsymbol{\beta}_1' \mathbf{t}_1 + \boldsymbol{\beta}_0' \mathbf{t}_0}}{\sum_u c(\mathbf{u}_1, \mathbf{t}_0) e^{\boldsymbol{\beta}_1' \mathbf{u}_1 + \boldsymbol{\beta}_0' \mathbf{t}_0}} = \frac{c(\mathbf{t}) e^{\boldsymbol{\beta}_1' \mathbf{t}_1}}{\sum_u c(\mathbf{u}_1, \mathbf{t}_0) e^{\boldsymbol{\beta}_1' \mathbf{u}_1}}. \qquad /14/$$

Mint látható, a *„nuisance"* paramétereket eliminálták a feltételes likelihoodból, az $L\left(\mathbf{t}_1 \mid \boldsymbol{\beta}_1'\right)$ feltételes permutációs eloszlás ismeretében pedig egzakt módon következtethetünk a $\boldsymbol{\beta}_1'$ paraméterekre, ami végül a $c(\mathbf{t})$ gyakoriságok generálását igényli. Ezt szolgálja az ún. *multivariate shift* algoritmus.

3.2. A „multivariate shift" algoritmus

Az egzakt feltételes következtetés alapja annak számszerűsítése, hogy az összes lehetséges 2^n számú **y** elrendezés tükrében az adott mintabeli szekvencia milyen eséllyel következik be. Egyféle megoldás generálni valamennyi olyan **y** vektort, melyekre $\mathbf{X}_0\mathbf{y}=\mathbf{t}_0$, és összeszámolni mindazon **y** vektorok számát, melyekre $\mathbf{X}_1\mathbf{y}=\mathbf{t}_1$ adódik.

A feladat méreteinek érzékeltetésére, tekintsünk egy háromváltozós (y, x_0, x_1) adatállományt, és keressük x_1 elégséges statisztikájának egzakt eloszlását az x_0 változó elégséges statisztikájának adottsága mellett.

3. tábla

Illusztratív adatok

Megfigyelés (*i*)	**y**	$\mathbf{x}_0$	$\mathbf{x}_1$
1	0	1	1
2	1	1	1
3	0	1	2
4	1	1	0

Most a mintabeli szekvencia **y**=(0,1,0,1)', $\mathbf{X}_0$=(1,1,1,1)' és $\mathbf{X}_1$=(1,1,2,0)'. Ezért az elégséges statisztikák vektora: **t**=(t_0,t_1)=[0(1,1)+1(1,1)+0(1,2)+1(1,0)]=(2,1). Így t_1 permutációs eloszlását keressük a t_0=2 feltétel mellett. Foglaljuk táblába a lehetséges 16 **y** vektort és a hozzájuk tartozó (t_0,t_1) értékeket:

4. tábla

A teljes mintatér: valamennyi lehetséges **y** *vektor*

Mintatér	y_1	y_2	y_3	y_4	t_0	t_1
1	0	0	0	0	0	0
2	0	0	0	1	1	0
3	0	0	1	0	1	2
4	0	0	1	1	2	2
5	0	1	0	0	1	1
6	0	1	0	1	2	1
7	0	1	1	0	2	3
8	0	1	1	1	3	3
9	1	0	0	0	1	1
10	1	0	0	1	2	1
11	1	0	1	0	2	3
12	1	0	1	1	3	3
13	1	1	0	0	2	2
14	1	1	0	1	3	2
15	1	1	1	0	3	4
16	1	1	1	1	4	4

Képezzük most a különböző (t_0,t_1) vektorok, majd a (t_0=2,t_1) vektorok gyakorisági eloszlását, melyeket az 5. és a 6. táblák közölnek:

5. tábla

A különböző (t_0,t_1) vektorok gyakorisági eloszlása

t_0	t_1	Gyakoriság	Valószínűség
0	0	1	1/16
1	0	1	1/16
1	1	2	2/16
1	2	1	1/16
2	1	2	2/16
2	2	2	2/16
2	3	2	2/16
3	2	1	1/16
3	3	2	2/16
3	4	1	1/16
4	4	1	1/16
Összesen		*16*	*1*

6. tábla

A különböző (t_0=2,t_1) vektorok gyakorisági eloszlása

t_0	t_1	Gyakoriság	Valószínűség
2	1	2	2/6
2	2	2	2/6
2	3	2	2/6
Összesen		*6*	*1*

Látható, hogy a feltételes eloszlást a feltétel nélküliből származtatni kézenfekvő, de magasabb mintanagyság mellett nem ésszerű. Gyorsabb megoldást eredményez a *Hirji–Mehta–Patel* [1987] által javasolt „multivariate shift" algoritmus, amit a 1. ábra illusztrál. Az algoritmus az alábbi rekurzív formulára épül:

$$\mathbf{t}_{i+1} = \mathbf{t}_i + y_{i+1}\mathbf{x}_{i+1}\,.$$

Az ábra egy fadiagram, melynek sorszámozott szintjei a megfigyelések egymásutániságát jelzik, minden számpár egy t_0,t_1 páros értékét mutatja, míg a mindenkori baloldali ágakat y=0, a jobboldali ágakat pedig y=1 azonosítja. Ennek megfelelően a következő (t_0,t_1) értéket mindig aszerint növeljük meg 0-val vagy x_i-vel (0x vagy 1x) értékkel, hogy baloldali, avagy jobboldali ágon szerepel.

A következő észrevételek a számlálási algoritmus gyorsítását szolgálják.

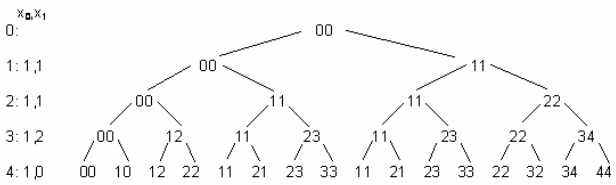
1. A második lépésben két (1,1) ág van mivel egymás után két azonos megfigyelés következik. E két (1,1) ág alatt azonos eredményekre jutunk, tehát az (1,1) ág alatti eredményeket vehetjük kétszeres gyakorisággal.

2. A 3. lépésben sem a (0,0) állapotból, sem a (3,t_1) állapotból nem tudunk egylépéses (1,2) hozzáadással (2,t_1) állapotba jutni. Ez a *megvalósíthatatlanság-kritérium* (*Hirji–Mehta–Patel* [1987]).

3. A megvalósíthatatlanság-kritérium annál hatékonyabbam működik, minél magasabb kovariánson kezdjük el előbb végrehajtani. Ha például példánkban a 4. x_0 érték 1 helyett 2 lenne, akkor a (0,0) állapotból rögtön (2,t_1) ágra kerülhetünk, ha ezzel kezdjük az eljárást.

4. Mivel az első két megfigyelés azonos kovariánsokkal bír, ezért a kombinálásukkal a 0. lépésről rögtön a második lépésre ugorhatunk úgy, hogy az induló (0,0) állapotot az i=0,1,2 csomópontokban i(1,1) értékkel növeljük, miközben a csomópontok gyakorisága $\binom{2}{i}$. Ezzel a keresési időt csökkentjük, de binomiális együtthatókat kell számítani.

7. ábra A „multivariate shift" algoritmus menete



3.3. Következtetés egyetlen paraméterre

Az egyedi β_1 paraméterre való következtetés a **T** változó azon *feltételes* eloszlásán alapul, mely csak a β_1 paraméter tekintetében változik, a többi paramétert pedig mint „zavaró" paramétert rögzíti:

$$L\left(t_1 \middle| \beta_1\right) = \frac{c(t_0, t_1, t_2, ..., t_p)e^{\beta_1 t_1}}{\sum_u c(t_0, u, t_2, t_2, ..., t_p)e^{\beta_1 u}}, \qquad /15/$$

ahol $c(t_0, u, t_2, ..., t_p) \geq 1$.

Az elégséges T_1-statisztika egzakt eloszlásának a használatát illusztrálja a következő kis esettanulmány. Egy 46 elemű véletlen minta struktúráját mutatja a 7. tábla, ahol 3 magyarázó változó 8 *különböző* kovariánsa magyarázza összesen f=29 darab y=1 előfordulását. A minta vállalkozásokat tartalmaz, melyekre y=1, ha felszámolási eljárás van ellene folyamatban (*csőd*), egyébként y=0, miközben a vállalkozás esetében x_1=1, ha az átlagosnál alacsonyabb a hosszú távú eladósodottsága, x_2=1, ha az átlagosnál jövedelmezőbb, és x_3=1, ha rövid távú likviditása az átlagosnál jobb, egyébként valamennyi másik x értéke zéró. A magyarázó változók $\mathbf{x}_k$ kovariánsa rendre n_k gyakorisággal fordul elő, melyből f_k számú y=1 tulajdonságú.

7. tábla

*Különböző kovariánsok megoszlása a mintában a „*csőd*“ gyakorisága szerint*

Elemszám		Kovariáns (**x**)		
n_k	f_k	x_1	x_2	x_3
3	3	0	0	0
2	2	0	0	1
4	4	0	1	0
1	1	0	1	1
5	5	1	0	0
5	3	1	0	1
9	5	1	1	0
17	6	1	1	1
t-statisztika	t_0=29	t_1=19	t_2=16	t_3=12

Vegyük észre, hogy az x_1 változó tekintetében a minta kváziszeparált, hiszen x_1=0 mellett nem fordul elő y=0 esemény. Következésképp a minta likelihoodja β_1 tekintetében nem maximálható. A mintában a β_j (j=0,1,2,3) paraméterek elégséges statisztikái /13/ felhasználásával rendre: t_0=f=29, t_1=19, t_2=16, t_3=12. Az Olvasó könnyen ellenőrizheti, hogy a t_1 elégséges statisztika megszorítás nélküli alsó határa t_1=19, felső határa pedig t_1=29, vagyis a minta t_1 tekintetében nem belső pontja az ún. konvex kiterjesztésnek.

A T_1 változó feltételes eloszlását jellemzendő, tekintsük a 8. táblát, mely a 7. tábla 29 *csőd* vállalkozásának egy olyan szekvenciában való elrendezését tartalmazza, mely megőrzi a [t_0=29, t_2=16, t_3=12] mintabeli értékeket, viszont a t_1 statisztika a t_1=26 értékre emelkedik. A 29 csőd vállalkozás természetesen sokféle szekvenciában elrendezhető, de mint arra a későbbiekben utalni fogunk, úgy nem, hogy a [t_0=29, t_2=16, t_3=12] feltétel mellett t_1 értéke magasabb legyen mint 26. Itt emlékeztetünk arra, hogy a [t_0=29, t_2=16, t_3=12] feltétel elhagyásával t_1 maximális értéke 29 volt.

8. tábla

Elégséges statisztikák a 29 csőd *eseménynek egy „alternatív" szekvenciája alapján*

Elemszám		Kovariáns (**x**)		
n_k	f_k	x_1	x_2	x_3
3	3	0	0	0
2	0	0	0	1
4	0	0	1	0
1	0	0	1	1
5	5	1	0	0
5	5	1	0	1
9	9	1	1	0
17	7	1	1	1
t-statisztika	t_0=29	t_1=26	t_2=16	t_3=12

Csődvizsgálatunkban a t_1 statisztika egzakt, feltételes eloszlását a 9. tábla közli. Mint látható, a [t_0=29, t_2=16, t_3=12] feltétel mellett nem található olyan szekvencia, mely kisebb t_1 értéket produkálna, mint 19, vagy nagyobbat, mint 26. Látható, hogy a konkrét minta t_1 terjedelmének a minimális értékéhez tartozik, és ez a **t** struktúra 29445360 különböző szekvencia esetén következik be. A **t** vektortól a csak a t_1=26 értékben különbözőt produkáló szekvenciák száma pedig 19448. Mint látható, az elégséges statisztika feltételes, permutációs eloszlásának a meghatározása számításigényes feladat, mely igen gyors algoritmust igényel. (Lásd *Trichler* [1984], *Hirji–Mehta–Patel* [1987], *Hirji* [1992], *Mehta–Patel–Senchaudhuri* [2000].) Az alkalmazott „multivariate shift" algoritmus lényegét a korábbiakban már tárgyaltuk.

9. tábla

A t_1-statisztika egzakt, feltételes eloszlása

t_1	$c(29,t_1,16,12)$
19	29,445,360
20	147,312,480
21	271,271,448
22	231,819,344
23	95,325,644
24	17,473,144
25	1,204,008
26	19,448
Összesen	*793,870,896*

A tábla gyakoriságait használva, például a t_1=19 esemény feltételes valószínűsége rögzített β_1 paraméter mellett:

$$L\left(t_1 = 19 \mid \beta_1\right) = \frac{c(29,19,16,12)e^{19\beta_1}}{\sum_{t_1=19}^{26} c(29,t_1,16,12)e^{t_1\beta_1}}.$$

Az elégséges statisztika feltételes eloszlását hipotézisvizsgálatra az alábbi módon használjuk.

Hipotézisek tesztelése

A parciális regressziós paraméterek tesztelése érdekében tekintsük az alábbi hipotézispárt:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0 .$$

Az egzakt *p*-értéket úgy nyerjük, hogy a /15/ valószínűség H_0 melletti értékeit összegezzük a specifikált **K** kritikus tartományon:

$$p = \sum_{v \in \mathbf{K}} L\left(v \mid \beta_1 = 0\right) .$$

Kritikus tartományt két alapvető módon képezhetünk. Egyfelől a *feltételes valószínűség*, másfelől a *feltételes score* elv alapján.

A feltételes valószínűség elvének megfelelően kritikus tartományt képeznek mindazon *v* értékek, melyekre a /15/ feltételes valószínűség nem nagyobb, mint a megfigyelt t_1-értékre számított feltételes valószínűség, vagyis: $L(v|0) \leq L(t_1|0)$. Így az egzakt *p*-érték:

$$p = \sum_{L(v|0) \leq L(t_1|0)} L\left(v \mid \beta_1 = 0\right) .$$

Mivel a nullhipotézis érvénye mellett $e^{t \cdot 0} = 1$, ezért a *p*-érték számítása a 9. tábla *c*(.) gyakoriságainak a megoszlásain alapszik. Esetünkben

$$p = L(19 \mid 0) + L(24 \mid 0) + L(25 \mid 0) + L(26 \mid 0) =$$

$$= \frac{29445360 + 17473144 + 1204008 + 19448}{793870896} = 0,061 .$$

Eszerint minden 6,1 százaléknál alacsonyabb szignifikanciaszinten elutasítjuk a nullhipotézist.

A feltételes score elv szerint viszont a kritikus tartományt azok a *v* értékek alkotják, melyekre

$$\left(\frac{v - \mu_1}{\sigma_1}\right)^2 \geq \left(\frac{t_1 - \mu_1}{\sigma_1}\right)^2 ,$$

ahol μ_1 és σ_1 a T_1 változó feltételes eloszlásának átlaga és szórása a β_1 paraméter zéró értéke mellett.

Paraméterbecslés

Célunk most $1-\alpha$ megbízhatóságú (β_a, β_f) konfidenciaintervallumot szerkeszteni a β paraméterre, mely definíció szerint eleget tesz a

$$Pr(\beta_a < \beta < \beta_f) = 1 - \alpha$$

követelménynek, ahol β_a a konfidenciaintervallum alsó, β_f pedig a felső határát jelöli, és $0<\alpha<1$ az alul- és a felülbecslés együttes kockázata. E két kockázatot egyenlően megosztva, majd a /15/ feltételes eloszlás kumulatív valószínűségeit képezve, a felső és az alsó határ definíció szerint rendre eleget tesz az alábbi azonosságoknak:

$$\sum_{v \le t_1} L\left(v \mid \beta_f\right) = \alpha / 2$$

$$\sum_{v \ge t_1} L\left(v \mid \beta_a\right) = \alpha / 2 .$$

Vegyük észre, hogy ha $t_1=t_{\max,}$ akkor (lévén teljes eseményrendszer) a kumulatív valószínűség 1, ezért *invariáns* β értékére, így ilyenkor megállapodás szerint $\beta_f=\infty$. Hasonlóan, ha $t_1=t_{\min}$, akkor $\beta_a=-\infty$.

A β_1 paraméter *pontbecslésére* kétféle lehetőség nyílik. Maximálhatjuk egyfelől β_1 tekintetében a /15/ szerinti $L(t_1|\beta_1)$ valószínűséget. Ez *egzakt feltételes maximum likelihood* β_{ECML} becslést eredményez, viszont nem működik, ha $t_1=t_{\min}$, vagy $t_1=t_{\max}$. Ebben az esetben alkalmazhatjuk a *torzítatlan medián* módszert, amely szerint:

$$\beta_{um} = (\beta_{f(.5)} + \beta_{a(.5)}) / 2 ,$$

ahol $\beta_{a(.5)}$ az $\alpha=0{,}5$ megbízhatóságú konfidenciaintervallum alsó, $\beta_{f(.5)}$ pedig a felső határa. Ha valamelyik határra végtelen adódna, akkor a pontbecslést automatikusan a másik határ jelenti. Esetünkben a β_1 paraméter 95 százalékos konfidencia tartományának felső határa az

$$L\left(19 \mid \beta_{1f}\right) = \frac{c(29,19,16,12)e^{19\beta_{1f}}}{\sum_{t_1=19}^{26} c(29,t_1,16,12)e^{t_1\beta_{1f}}} = 0,025$$

azonosság iteratív megoldásával $\beta_{1f}=0{,}16$ adódik.

Előrejelzés

Becsüljük az $\mathbf{x}_0$ kovariáns mellett az „y=1” esemény P_0 valószínűségének egzakt konfidencia intervallumát. Az előrejelzés érdekében paraméterezzük át a logit modellt az alábbi módon:

$$\operatorname{logit}\left(\pi_x\right) = \left(\beta_0 + \mathbf{x}_0'\boldsymbol{\beta}\right) + \left(\mathbf{x}' - \mathbf{x}_0'\right)\boldsymbol{\beta} . \qquad /16/$$

A /16/ modell $\left(\beta_0\right)^* = \beta_0 + \mathbf{x}_0'\boldsymbol{\beta}$ tengelymetszetének egzakt konfidenciaintervalluma egyben konfidenciaintervallum az $\ln\left(P_0/(1-P_0)\right)$ logitra. A modell új magyarázóváltozói

az eredeti értékeknek az előrejelzési ponttal csökkentett értékei, valamennyi mintaelemre. A logitra nyert konfidencia határok végül a megszokott exp(.)/(1+exp(.)) módon adják a valószínűségre vonatkozó határokat.

Illusztratív számítások

Jelen példában a paraméterbecslés eredményeit a 10. tábla tartalmazza.

10. tábla

Egzakt következtetés a paraméterekre

Paraméter	Pontbecslés	Egzakt 95% CI	Egzakt *p*-érték
β_0	3,535	1,477 – (∞)	0,0001
(β_0)*	–0,737	–1,910 – (0,310)	0,164
β_1	–1,886	–∞ – (0,160)	0,061
β_2	–1,548	–4025 – (0,363)	0,117
β_3	–1,156	–2,997 – (0,512)	0,154

A regressziós meredekségeket illetően, a *p*-értékek alapján látszik, hogy míg az x_1 változó 6,1 százalékos szinten szignifikáns, addig x_2 és x_3 nem. Mivel a β_0 tengelymetszet a (0,0,0) kovariáns melletti lineáris előrejelzés, ezért egy vállalkozás esetében, ha jelentősen eladósodott, kevéssé jövedelmező, és kevéssé likvid, akkor a „*csőd*" bekövetkezésének valószínűsége 95 százalékos megbízhatósággal *legalább*

$$\frac{e^{1,48}}{1+e^{1,48}} = 0,814 .$$

Számításainkban a β_0^* tengelymetszetet oly módon becsültük, hogy a magyarázóváltozók valamennyi értékéből egyöntetűen 1-et levontunk. Ezáltal β_0^* az $\mathbf{x}_0$=(1,1,1) kovariáns (kevéssé eladósodott, igen jövedelmező és módfelett likvid) mellett becsült logit, így e kovariáns mellett a „*csőd*" bekövetkezésének esélye *legalább*

$$\frac{e^{-1,91}}{1+e^{-1,91}} = 0,129$$

és legfeljebb

$$\frac{e^{0,31}}{1+e^{0,31}} = 0,577 .$$

Másfelől, a β_1 regressziós meredekség parciális értelmezését illetően, ha a vállalkozás kevéssé eladósodott, akkor *ceteris paribus*

$$e^{0,16} = 1,173 ,$$

tehát 17,3 százalékkal nagyobb a csődhelyzetbe kerülés odds-aránya szemben azokkal, akik inkább eladósodottak. Ugyanakkor 95 százalékos megbízhatósági szinten legfeljebb

$$\frac{e^{0,16}}{1+e^{0,16}} = 0,54$$

a valószínűsége annak, hogy a szóban forgó vállalkozás csődhelyzetbe kerül. Vegyük észre, hogy ez egy *legfeljebb* jellegű becslés, hiszen a *negatív előjelű megfelelő pontbecslést* alkalmazva az analóg eredmények rendre

$$e^{-1,886} = 0,152$$

$$\frac{e^{-1,886}}{1+e^{-1,886}} = 0,132,$$

amely valószínűségek a csődhelyzetbe kerülés valószínűségének a csökkenésére utalnak azokkal szemben, akik az átlagosnál kevésbé eladósodottak. A β_2 és β_3 paraméterekre való következtetés hasonló módon történik.

3.4. Következtetés több paraméterre

A hipotézisek tesztelésének megkülönböztetett esete a regressziós paraméterek vonatkozásában, amikor a hipotézis tartalmilag egy kvalitatív magyarázó változóra vonatkozik, melynek lehetséges kimenetei kategóriák. Ekkor e kategóriaváltozót (a kategóriák számától függően) bináris, ún. *dummy* változók rendszerével írjuk le, tehát e dummy változók paramétereit kell tesztelnünk. Ilyenkor a statisztikai hipotézis szükségszerűen a paraméterek egy *csoportjára* vonatkozik egyidejűleg.

Egy másik esettanulmányra áttérve, csődhelyzet szempontból tekintsünk 47 vállalkozást *likviditási rátájuk*[4], és *bonitásuk*[5] szempontjából, és bontsuk mindkét mutató skáláját három, rendre: *alacsony*, *átlagos*, és *magas* kategóriára. Legyen x_1=1, ha a likviditási ráta alacsonyabb az átlagosnál, x_2=1, ha a likviditás átlagos szintű, x_3=1, ha a bonitás alacsonyabb az átlagosnál, és x_4=1, ha átlagos színvonalú a bonitás, egyébként minden más x érték zéró. Ha a vállalkozás csődhelyzetben van, akkor y=1, egyébként y=0. Bár a modellben csupán két tényezőt, nevezetesen a *likviditást* és a *bonitást* vizsgáljuk, módszertanilag több, rendre 2-2 dummy jellegű (1,0 kimenetű) x magyarázóváltozó reprezentálja e tényezőket. Az adatokat a 11. tábla közli. A táblában feltüntetésre került a tengelymetszethez tartozó mesterséges $\mathbf{x}_0$ összegző vektor, továbbá az elégséges t-statisztikák értéke is.

[4] Likviditási ráta = (forgóeszközök – készletek) / rövid lejáratú kötelezettségek.
[5] Bonitás = hosszú lejáratú kötelezettségek / saját vagyon.

11. tábla

A likviditás és a bonitás hatása a csődhelyzetbe kerülésre

Csődhelyzet (y)	Tengelymetszet	Kovariáns				Gyakoriság
	$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	
1	1	1	0	0	0	1
1	1	0	1	0	0	2
1	1	1	0	1	0	4
1	1	0	1	0	1	4
1	1	0	0	0	0	1
1	1	0	1	1	0	2
0	1	1	0	1	0	3
0	1	0	1	0	1	8
0	1	0	0	0	0	2
0	1	0	1	1	0	5
0	1	0	0	1	0	2
0	1	0	0	0	1	13
t-statisztika	14	5	8	6	4	–

A logit modell ezek után

$$\ln\left(odds_x\right)=\text{logit}\left(\pi_x\right)=\beta_0+\beta_1 x_1+\beta_1 x_2+\beta_3 x_3+\beta_4 x_4\,,$$

amely paramétereit a fenti minta alapján a maximum likelihood módszerrel nem tudjuk becsülni, mivel a megfigyelt adatok a mintatér határára esnek. Mindazonáltal érdekes számunkra, hogy a bonitási ráta szignifikánsan befolyásolja-e a csődhelyzetet, vagy sem. Ezt formailag a

$$H_0:\beta_3=\beta_4=0 \qquad /17/$$

hipotézis fejezi ki, melyet a megfelelő elégséges statisztikák egzakt, együttes feltételes eloszlására támaszkodva tesztelhetünk.

Általánosságban tekintsük a $\boldsymbol{\beta}$ paramétervektornak az első q-elemét leválasztó $\boldsymbol{\beta}_0^{'}$ és a további $(p–q)$-elemét tartalmazó $\boldsymbol{\beta}_1^{'}$ partícióját, melyekhez rendre az elégséges statisztikák $\mathbf{t}_0$ és $\mathbf{t}_1$ vektora tartozik. Ekkor a paraméterek $\boldsymbol{\beta}_1^{'}$ csoportjára való *egyidejű* következtetés a hozzájuk tartozó $\mathbf{T}_1$ elégséges statisztikák *együttes* eloszlásának az ismeretét igényli, a többi elégséges $\mathbf{T}_0=\mathbf{t}_0$ statisztika mintabeli szinten való rögzítettsége mellett. Az így definiált feltételes eloszlás tehát független a $\boldsymbol{\beta}_0^{'}$ paraméterektől.

Legyen nullhipotézisünk, hogy valamenyi $\boldsymbol{\beta}_1^{'}$ paraméter zéró:

$$H_0:\boldsymbol{\beta}_1^{'}=\mathbf{0}^{'}$$

szemben a kétoldali H_1 alternatívával, miszerint $\boldsymbol{\beta}_1^{'}$ elemei közül legalább egy nem zéró. A hipotézis tesztelésére szolgáló feltételes valószínűség:

$$L\left(\mathbf{t}_1 \mid \boldsymbol{\beta}_1'\right) = \frac{c\left(\mathbf{t}_1, \mathbf{t}_0\right) e^{\boldsymbol{\beta}_1' \mathbf{t}_1}}{\sum_{\mathbf{u}} c(\mathbf{u}_1, \mathbf{t}_0) e^{\boldsymbol{\beta}_1' \mathbf{u}}}. \qquad /18/$$

Az egzakt kétoldali *p*-érték általánosságban

$$p = \sum_{\mathbf{v} \in \mathbf{K}} L\left(\mathbf{v} \mid \boldsymbol{\beta}_1' = \mathbf{0}\right).$$

A kritikus tartományt most is kétféle úton konstruálhatjuk. A feltételes valószínűség módszerével mindazon **v** értékek alkotják a **K** kritikus tartományt, melyekre:

$$\mathbf{K}_{prob} = \left\{\mathbf{v} : L\left(\mathbf{v} \mid \boldsymbol{\beta}_1' = \mathbf{0}\right) \le L\left(\mathbf{t}_1 \mid \boldsymbol{\beta}_1' = \mathbf{0}\right)\right\}, \qquad /19/$$

míg a feltételes score elv alapján

$$\mathbf{K}_{score} = \left\{\mathbf{v} : \left(\mathbf{v} - \boldsymbol{\mu}_1\right)' \boldsymbol{\Sigma}_1^{-1} \left(\mathbf{v} - \boldsymbol{\mu}_1\right) \ge \left(\mathbf{t}_1 - \boldsymbol{\mu}_1\right)' \boldsymbol{\Sigma}_1^{-1} \left(\mathbf{t}_1 - \boldsymbol{\mu}_1\right)\right\}, \qquad /20/$$

ahol $\boldsymbol{\mu}_1$ az átlaga, $\boldsymbol{\Sigma}_1$ pedig a kovarianciamátrixa az $L\left(\mathbf{t}_1 \mid \boldsymbol{\beta}_1'\right)$ eloszlásnak.

Vizsgálatunkban a /17/ hipotézis tesztelése az $L(t_3,t_4|\beta_3=\beta_4=0)$ permutációs null-eloszlás alapján történik, a többi elégséges statisztika megfigyelt (t_0=14, t_1=5, t_2=8) értéken való rögzítése mellett. A teszteléshez az egzakt feltételes *score* módszert használjuk. A feltételes *score* értéke most általában

$$s = \left[t_3 - \mu_3, t_4 - \mu_4\right] \begin{bmatrix} \mathrm{var}(t_3) & \mathrm{cov}(t_3, t_4) \\ \mathrm{cov}(t_3, t_4) & \mathrm{var}(t_4) \end{bmatrix}^{-1} \begin{bmatrix} t_3 - \mu_3 \\ t_4 - \mu_4 \end{bmatrix},$$

ahol esetünkben t_3=6 és t_4=4. Így – az átlag és a kovarianciamátrix közlésétől eltekintve –, s=7,293.

A következtetés alapjául szolgáló *p*-érték meghatározása innen kétirányú. Egyfelől a feltételes score teszt alapján a /20/ kritikus tartományt alkalmazva a *p*-érték

$$p_{score} = \sum_{t_3, t_4 \in \mathbf{K}} L\left(t_3, t_4 \mid \beta_3 = \beta_4 = 0\right) = 0,0256.$$

Másfelől aszimptotikus *p*-érték meghatározására is lehetőség nyílik a 2 szabadságfokú chi-négyzet eloszlás tengelyén az alábbi szárny-valószínűség (*tail-probability*) számítása révén:

$$\Pr\left(\chi_2^2 > 7,293\right) = 0,0261.$$

A likviditásra vonatkozó analóg egzakt, illetve aszimptotikus *p*-értékek rendre: 0,007 és 0,009. Mint látható, a likviditás inkább szignifikáns tényező mint a bonitás. Ugyanak-

kor jelen esetben, a kicsi mintaelemszám ellenére az egzakt és az aszimptotikus eredmények nagyon közel állnak egymáshoz. Az aszimptotikus eredmények pontosságának most az a magyarázata, hogy nem a feltétel nélküli, hanem a feltételes ML-becsléshez kötődnek. Felhívjuk a figyelmet újra, hogy bár a paraméterek nem egzakt maximum likelihood becslésére nincs lehetőség a vizsgált adatállomány esetén, egzakt módon lehetőség nyílt hipotézisek tesztelésére.

3.5. Következtetés rétegzett minta esetén

Tekintsünk egy g=1,2,...,m számú rétegre bontott sokaságot, amely rétegekből rendre n_1, n_2,...,n_m elemű független minták állnak rendelkezésre, melyekben rendre f_1, f_2,...,f_m az y=1 esetek száma. Jelölje π_{ig} a $\Pr(Y_{ig}=1|\mathbf{x}_{ig})$ esemény feltételes valószínűségét, ahol $\mathbf{x}_{ig}$ a p-dimenziós kovariáns a g rétegben, az i egyedre vonatkozóan. E körülmények között a logit az

$$\ln\left(\frac{\pi_{ig}}{1-\pi_{ig}}\right)=\beta_{0g}+\beta_1 x_{ig1}+\beta_2 x_{ig2}+\ldots+\beta_p x_{igp}$$

lineáris modell szerint alakul, ahol a β_j parciális meredekség közös valamennyi rétegre, és a β_{0g} rétegspecifikus tengelymetszet fejezi ki a réteghatást (a réteghez való tartozást *dummy* változók rendszere rögzíti az adatok között). E körülmények között az elégséges statisztikák képzése

$$t_j=\sum_{g=1}^{m}\sum_{i=1}^{n_g} y_{ig}x_{igj} \qquad /21/$$

módon történik, ahol $t_{0g}=f_g$ minden rétegre. A réteghatások becslése nagymértékben növeli a becsülendő paraméterek számát, ezért, ha nem célunk a réteghatások elemzése, akkor kézenfekvő „zavaró" paraméterként kezelni azokat, és elégséges statisztikáik megkötése mellett következtetni a réteghatástól mentes többi paraméterre.

Az alábbiakban 30 vállalkozást tekintünk, akik tevékenységük alapján hét adott ágazat egyikéhez kötődnek, alakuláskori becsült kockázati indexük a K=0,1,2,...,15 skála valamely kimenete, és y=1, ha 3 éven belül fizetésképtelenség bejelentése miatt indult ellenük csődeljárás, egyébként y=0. A rétegzés ágazatok szerint (g=1,2,3,4,5,6,7) történt. Az adatok a 12. táblában láthatók.

A csődeljárás π_{ig} valószínűségi modellje most a K kockázati index feltétele mellett a következő:

$$\ln\left(\frac{\pi_{ig}}{1-\pi_{ig}}\right)=\beta_{0g}+\beta K_{ig}\ .$$

12. tábla

Az alakuláskori kockázati index hatása a csődeljárás megindítására

Vállalkozás	Ágazat (*g*)	Kockázati index (*K*)	Csődeljárás (*y*)
1	1	0	0
2	3	9	1
3	4	2	1
4	1	3	0
5	2	2	1
6	5	6	0
7	1	2	0
8	6	3	0
9	7	2	0
10	6	0	1
11	1	7	0
12	6	0	0
13	4	0	0
14	1	15	1
15	2	0	0
16	1	3	0
17	3	2	0
18	4	0	0
19	1	2	0
20	3	1	0
21	4	0	0
22	1	6	0
23	7	6	1
24	1	5	0
25	6	0	0
26	4	0	0
27	5	3	1
28	1	2	0
29	6	0	0
30	5	0	1

A kevés megfigyelés miatt, ha a közös kockázati tényező becslése mellett az ágazati hatásokat is becsülnénk, akkor ez a paraméterek relatíve magas száma miatt nagymértékben rontaná a kockázati index hatásának a becslését. Ezért a β_{0g} ágazati hatás becslését eliminájuk azáltal, hogy az ágazatonkénti csődeljárások számát feltételként kezelve következtetünk a β paraméterre:

$$H_0 : \beta = 0, \quad H_1 : \beta \neq 0 .$$

Ezt most megtehetjük egzakt és aszimptotikus úton is. Az egzakt, feltételes eloszláson alapuló maximum likelihood becslés β_{ECML}=0.325, a hozzá tartozó (feltételes score) *p*-érték 0,0167, míg a 95 százalékos konfidenciaintervallum [0,0223, 0,741]. Ugyanakkor a feltételes, de aszimptotikus módon becsült 95 százalékos megbízhatóságú konfidenciaintervallum [–0,004, 0,654]. A megfelelő *p*-értékek pedig rendre: *p*(score)=0,0129, *p*(Wald)=0,0528, és *p*(likelihood ratio) = 0,023.

Látható, hogy az egzakt módszert alkalmazva a kockázati tényező hatása (mind a konfidenciaintervallum, mind a *p*-érték alapján) 5 százalékos szinten szignifikáns. Ezzel szemben az aszimptotikus konfidenciaintervallum és az aszimptotikus Wald-teszt alapján a kockázati index hatása nem szignifikáns.

3.6. A paraméterek lineáris kombinációjának tesztelése

Tekintsük végül a logit lineáris regresszióját az alábbi formában

$$\text{logit}(.) = \mathbf{X}_{(n,p)}\boldsymbol{\beta}_{(p,1)}, \qquad /22/$$

ahol $\mathbf{X}$ az adatmátrix és a $\boldsymbol{\beta}$ vektor tartalmazza a tengelymetszetet is. Legyen feladatunk a

$$H_0 : \mathbf{C}_{(r,p)}\boldsymbol{\beta}_{(p,1)} = \mathbf{0} \qquad /23/$$

hipotézis tesztelése, ahol $\mathbf{C}$ rangja teljes. H_0 tesztelése érdekében írjuk fel a /22/ modellt az alábbi átparaméterezett formában

$$\text{logit}(.) = \mathbf{X}\boldsymbol{\beta} + \left(\left(\underbrace{\mathbf{XG}}_{\mathbf{X}_2} \right) \left(\underbrace{\mathbf{C}\boldsymbol{\beta}}_{\mathbf{b}_2} \right) \right) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 ,$$

ahol a $\mathbf{G}$ mátrixot úgy választjuk meg, hogy $\mathbf{GC}=\mathbf{0}$ teljesüljön, és $\mathbf{X}_1=\mathbf{X}$, valamint $\boldsymbol{\beta}_1=\boldsymbol{\beta}$. Így a /23/ hipotézis tesztelése az $L(\mathbf{T}_2|\mathbf{T}_1=\mathbf{t}_1)$ egzakt eloszlás meghatározásával végrehajtható.

4. TORZÍTÁSCSÖKKENTŐ KORREKCIÓK ASZIMPTOTIKUS KÖVETKEZTETÉSEKHEZ

Ha az „1” megfigyelés *ritka esemény* a mintában, akkor további „1” esetek csatolása kívánatos a mintához, standard hiba csökkentő hatása révén. Ha ugyanis a logit modell előrejelzése megbízható, akkor $\pi_i|y_i=1$ becsült értéke magasabb mint $\pi_i|y_i=0$ becsült értéke, de 0,5-höz közeli, mert alulbecsült, tehát $\pi_i(1-\pi_i)$ értéke a /11/ formulában relatíve magas, és nagyobb az „1”, mint a „0” egyedek esetén, tehát újabb „1” egyed csatolása a mintához a paraméterek varianciáját tovább csökkenti. Ha az „1” ritkasága miatt csatolására nincs lehetőség, akkor célszerű alkalmas módon „0” egyedeket elhagyni (*King–Zeng* [2001a,b]). Ennek megfelelő mintavételi stratégia az ún. *case-control* módszer, ahol adott kategóriához tartozó „*csőd*” esethez választunk egy vagy több „0”, azaz „*kontroll*” jellegű megfelelő vállalkozást. A *case* és *control* megfigyelések közel egyenlő részaránya a mintában az optimális arány a paraméterek standard hibája szempontjából. Az ilyen jellegű becslés további korrekciót, nevezetesen *prior* korrekciót igényel, ha van ilyen információnk az „1” egyedek sokasági P arányára vonatkozóan. A csődbement vállalkozásokra ilyen jellegű információ rendelkezésre áll.

A *torzításcsökkentő prior korrekció* módszere (*Prentice–Pyke* [1979], *Manski–Lerman* [1977]) a klasszikus ML-becslésből indul ki, majd a becsléseket korrigálja az y=1 egyedek *a priori* sokasági P arányára, és a mintabeli $\bar{y}$ arányára vonatkozó információval. A tengelymetszet konzisztens korrigált becslése:

$$\hat{\beta}_0 - \ln\left[\left(\frac{1-P}{P}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] = \hat{\beta}_0 - \left[\underbrace{\ln\left(\frac{\bar{y}}{1-\bar{y}}\right)}_{\hat{\beta}_{0ML}} - \underbrace{\ln\left(\frac{P}{1-P}\right)}_{\beta_0}\right]$$

Ennek mondanivalója, hogy ha valamennyi magyarázóváltozó értéke zéró, akkor ez az odds-arány ismert, mégpedig $P/(1–P)$. A logit modell által becsült odds-arány $e^{\hat{\beta}_0}$, viszont a ML odds-arány $\bar{y}/(1-\bar{y})$. A korrekció a becsült tengelymetszetet a torzítás mértékével módosítja, és hatására az **x=0** nevezetes esetben a modell által becsült odds-arány a sokasági odds-arányt adja.

Az elemzések többségében a hangsúly nem föltétlenül a regressziós paraméterek becsült értékének az elemzésén, hanem a valószínűségek minél pontosabb számításán van. Ilyenkor mind a tengelymetszet, mind a regressziós paraméterek minél precízebb becslése központi kérdés, melynek egyféle eszköze a prior korrekció módszere. Hátránya a prior korrekció módszerének, hogy ha a modell tévesen specifikált, akkor a becslések kevésbé robusztusak (lásd *Xie–Manski* [1989]), mint az alább tárgyalandó módszer.

A súlyozott mintavételi maximum likelihood becslés (Manski–Lerman [1977]) egy alternatív módszer az „1” tulajdonságú egyedek eltérő sokasági és mintabeli arányának a figyelembe vételére, ahol $\Pr(Y=1\,|\,\mathbf{x}) = \pi_x^{v_1}$ és $\Pr(Y=0\,|\,\mathbf{x}) = (1-\pi_x)^{v_0}$ definiálja a feltételes valószínűségeket. A nem csoportosított i=1,2,...,n minta esetén ekkor a likelihood függvény

$$L_v = \prod_{i=1}^{n}\left(\pi_i^{v_1}\right)^{y_i}\left(\left(1-\pi_i\right)^{v_0}\right)^{1-y_i},$$

ahol

$$v_1 = \frac{P}{\bar{y}}, \quad v_0 = \frac{1-P}{1-\bar{y}}.$$

Látható, hogy ha a sokasági és a mintabeli arányok megegyeznek, akkor a klasszikus likelihood függvényt kapjuk. Ha $P > \bar{y}$ akkor csökkentjük a π_i valószínűség hatását a likelihoodban, egyébként növeljük. Mivel általában a súlyozatlan loglikelihood az

$$\ln L = -\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{\beta}'\mathbf{x}_i}\right) \qquad /24/$$

formában is írható, a maximálandó súlyozott loglikelihood függvény /24/ alapján egy tagban fölírva

$$\ln L_v = -\sum_{i=1}^{n}\left(v_1 y_i + v_0\left(1-y_i\right)\right)\ln\left(1+e^{(1-2y_i)\boldsymbol{\beta}'\mathbf{x}_i}\right) = -\sum_{i=1}^{n} v_i \ln\left(1+e^{(1-2y_i)\boldsymbol{\beta}'\mathbf{x}_i}\right).$$

A fenti felírás gyakorlati haszna az, hogy a v_i súlyokat meghatározva a paraméterek becslése bármely standard „*logistic regression"* programmal számítható. A módszer hiányossága, hogy a megszokott információs mátrixon alapuló standardhiba-számítás erősen torzított becslést eredményez, másfelől a ritka esemény mintán belüli ritkaságát (prior korrekció nélkül) nem veszi figyelembe. E hiányosságok kiküszöbölését teszik lehetővé a következő (*King–Zeng* [2001a]) korrekciók. A közelítőleg torzításmentes becslés érdekében végrehajtandó korrekció

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - bias\left(\hat{\boldsymbol{\beta}}\right),$$

ahol a torzítás mértékét /12/ szerint határozzuk meg[6]

$$\xi = 0{,}5\left(\left(1+v_1\right)\hat{\pi}_i - v_1\right)\left[\mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\right]_{ii}$$
$$\mathbf{W} = diag\left[\hat{\pi}_i\left(1-\hat{\pi}_i\right)v_i\right]$$

A fenti eljárás standardhiba-csökkentő hatású, mivel *McCullagh* és *Nelder* ([1989] 457. old.) alapján közelítőleg

$$\tilde{\boldsymbol{\beta}} \approx \frac{n}{n+p+1}\hat{\boldsymbol{\beta}},$$

ahol $n/(n+p+1) < 1$, és így

$$\mathbf{C}_{\tilde{\beta}} \approx \left(\frac{n}{n+p+1}\right)^2 \mathbf{C}_{\hat{\beta}}.$$

Természetesen a súlyozott likelihood maximálása, és a prior korrekció együtt is alkalmazható.

Ezen a ponton merül fel a *ritkaság* problémája, miszerint az „1" esemény mintabeli ritkasága miatt – bár $\tilde{\boldsymbol{\beta}}$ már közel torzítatlan –, a $\tilde{\pi}_i\left(\tilde{\boldsymbol{\beta}}\right)$ valószínűség alulbecsli a π_i valószínűséget. Ezt a faktort veszi figyelembe a feltételes valószínűség pontbecslésekor a *ritkasági* korrekció, mely a π_i valószínűséget *bayesi* szemléletben mint várható értéket definiálja (rögzített $\mathbf{x}_0$ kovariáns mellett):

[6] Most $\mu_i = \pi_i^{v_1}$, $\mu_i' = v_1\pi_i^{v_1}\left(1-\pi_i\right)$, $\mu_i'' = v_1\pi_i^{v_1}\left(1-\pi_i\right)\left(v_1 - \left(1+v_1\right)\pi_i\right)$.

$$\Pr\left(Y_0 = 1\right) = \pi_0 = E\left\{\frac{1}{1+e^{-\boldsymbol{\beta}'\mathbf{x}_0}}\right\}. \qquad /25/$$

A /25/ várható érték közelítő meghatározása érdekében képezzük az $1/\left(1+e^{-\boldsymbol{\beta}'\mathbf{x}_0}\right)$ függvény Taylor-sorát a $\tilde{\boldsymbol{\beta}}$ becslés körül, mely a kvadratikus taggal bezárólag:

$$\Pr\left(Y_0 = 1\right) = \tilde{\pi}_0 + \left[\frac{\partial \pi_0}{\partial \boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right) + \frac{1}{2}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)'\left[\frac{\partial^2 \pi_0}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right).$$

A várható értéket véve végül (a szükséges átalakításokat lásd *King–Zeng* [2001a]):

$$\pi_0 \approx \tilde{\pi}_0 + \left(0,5-\tilde{\pi}_0\right)\tilde{\pi}_0\left(1-\tilde{\pi}_0\right)\mathbf{x}_0'\mathbf{C}_{\tilde{\beta}}\mathbf{x}_0$$

adódik. Látható, hogy ha $0,5 > \tilde{\pi}_0$, és a logit paraméterek mintavételi kovarianciamátrixa nem zéró mátrix, akkor $\tilde{\pi}_0$ alulbecsli a π_0 valószínűséget.

IRODALOM

ALBERT, A. – ANDERSON, J. A. [1984]: On the existence of maximum likelihood estimates in logistic models. *Biometrica.* 71. évf. 1–10. old.

Bartus T. [2003]: Logisztikus regresszós eredmények értelmezése. *Statisztikai Szemle*. 81. évf. 4. sz. 328–347. old.

BRESLOW, N. E. – DAY, N. E. [1980]: Statistical Methods in Cancer Research. IARC. Lyon.

BULL, SB. – MAK, C. – GREENWOOD, C. M. T. [2002]: A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis*. 39. évf. 57–74. old.

CHRISTMANN, A. [2002]: Classification based on the support vector machine and on regression depth. In: *Dodge, Y.* (szerk.) *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Series: Statistics for industry and technology. Birkhaeuser. Basel. 341–352. old.

CHRISTMANN, A. – FISCHER, P. – JOACHIMS, T. [2002]: Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*. 17. évf. 273–287. old.

CHRISTMANN, A. – ROUSSEEUW, P. J. [2001]: Measuring overlap in logistic regression. *Computational Statistics and Data Analysis*. 37. évf. 65–75. old.

COLLETT, D. [1999]: *Modelling Binary Data*. Boca Raton. FL: CRC Press.

COX, D. R. – SNELL, E. J. [1989]: *Analysis of Binary Data*. Chapman and Hall. London.

CRAMER, J. S. [1999]: Predictive Performance of the Binary Logit Model in Unbalanced Samples. *The Statistician*. 48. évf. 85–94. old.

FONG, A. P. – YU, Y. H. – HEISEY, D. M. [1999]: Logistic Regression in an Adaptive Web Cache. *IEEE Internet Computing*. 3. sz. 27–36. old.

GARTHWAITE, P. H. – JOLLIFFE, I. T. – JONES, B. [1995]: *Statistical Inference*. Prentice Hall.

HAJDU, O.– VIRÁG, M. [2001]: A Hungarian Model for Predicting Financial Bankruptcy. *Society and Economy*. XXIII. évf. 1–2. sz. 28–46. old.

HIRJI, K. F. [1992]: Exact distributions for polytomous data. *JASA*. 87. évf. 487–492. old.

HIRJI,K. F. – MEHTA, C. R. – PATEL, N. R. [1987]: Computing distributions for exact logistic regression. *JASA*. 82. évf. 1110–1117. old.

HIRJI, K. F. – MEHTA, C. R. – PATEL, N. R. [1988]: Exact inference for matched case-control studies. *Biometrics*. 44. évf. 803–814. OLD.

HIRJI, K. F. – TSIATIS, A. A. – MEHTA, C. R. [1989]: Median unbiased estimation for binary data. *The American Statistician*. 43. évf. 7–11. old.

Hunyadi L. [2001]: Statisztikai következtetéselmélet közgazdászoknak. Központi Statisztikai Hivatal. Budapest.

JENNRICH, R. I. – MOORE, R. H. [1975]: Maximum Likelihood Estimation by Means of Nonlinear Least Squares. Proceedings of the Statistical Computing Section. *American Statistical Association*. 57–65. old.

KING, G. – ZENG, L. [2001a]: Logistic Regression in Rare Events Data. *Political Analysis*. 9. sz. 137–163. old.

KING, E. N. – RYAN, T. P. [2002]: A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression. *The American Statistician*. 56. évf. 3. sz. 163–170. old.

KING,G. – ZENG, L. [2001b]: Explaining Rare Events in International Relations. *International Organization*. 55. évf. 693–715. old.
MANSKI, CHARLES F. – LERMAN, STEVEN R. [1977]: The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica.* 45. évf. 8. sz. 1977–1988. old.
MEHTA, C. R. – PATEL, N. R. [1995]: Exact Logistic Regression: Theory and Examples. *Statistics in Medicine*. 14. évf. 2143–2160. old.
MEHTA, C. R. – PATEL, N. R. – SENCHAUDHURI, P. [2000]: Efficient Monte Carlo Methods for Conditional Logistic Regression. *JASA.* 95. évf. 449. sz. Theory and Methods. 99–108. old.
MCCULLAGH, P. – NELDER, J. A. [1989]: *Generalized Linear Models*. Chapman and Hall. New York.
PRENTICE, R. L. – PYKE, R. [1979] Logistic Disease Incidence Models and Case-Control Studies. *Biometrica.* 66. évf. 403–411. old.
SANTNER, T. J. – DUFFY, D. E. [1986]: A Note on A.Albert's and J.A.Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrica*. 73. évf. 755–758. old.
SCHAEFER, R. L. [1983]: Bias Correction in Maximum Likelihood Logistic Regression. Statistics in Medicine. 2. sz. 71–78. old.
TRICHLER, D. [1984]: An Algorithm for Exact Logistic Regression. *JASA.* 79. évf. 709–711. old.
XIE, YU – MANSKI, C. F. [1989]: The Logit Model and Response-Based Samples. *Sociological Methods and Research.* 17. évf. 3. sz. 283–302. old.

SUMMARY

The paper deals with the problems of inference for the logistic regression model caused by a small sample size. In fact, the small sample based inference is unavoidable when the research is about relatively rare events such as financial bankruptcy observed in special branches. The problems of interest are – on the one hand – that even provided a considerable sample size the customary unconditional asymptotic maximum likelihood estimation (UAML) does not exist when the sample is separated. On the other hand, in the case of an unbalanced sample the UAML estimator is biased to a great extent with no regard to the sample size. Fortunately, the so-called exact logistic regression is the appropriate tool for analysing such types of data. The paper discusses the underlying theory behind the exact conditional inference and provides illustrative examples – in the field of predicting financial bankruptcy – that contrast the exact inference with the more customary unconditional asymptotic maximum likelihood approach.