Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie
Metody analizy danych

923

ISSN 1898-6447
Zesz. Nauk. UEK, 2013; 923: 87–98
DOI: 10.15678/ZNUEK.2013.0923.07

Kamil Fijorek
Katedra Statystyki
Uniwersytet Ekonomiczny w Krakowie

Aproksymacja modelu regresji logistycznej Firtha za pomocą ważenia obserwacji

Streszczenie

W artykule przedstawiono model regresji logistycznej Firtha w kontekście wag przypisywanych przez metodę poszczególnym obserwacjom ze zbioru danych. Następnie dokonano przekształcenia modelu HLM do podobnej postaci. Na podstawie wniosków płynących z alternatywnego spojrzenia na model Firtha oraz HLM zaproponowano dwie metody aproksymacji modelu Firtha. Symulacyjnie zbadano jakość aproksymacji oraz omówiono praktyczne korzyści płynące z jej stosowania.

Słowa kluczowe: regresja logistyczna, redukcja obciążenia, całkowite rozdzielenie, estymacja przybliżona.

1. Wprowadzenie

Autor niniejszego artykułu porównał dwie metody estymacji modelu regresji logistycznej rozwiązujące problem całkowitego rozdzielenia (*complete separation*) [Fijorek 2012]. Wspomniany problem występuje, gdy sukcesy i porażki mogą być całkowicie rozdzielone za pomocą jednej zmiennej niezależnej lub liniowej kombinacji kilku zmiennych niezależnych. W takim przypadku zastosowanie estymacji metodą największej wiarygodności (MNW) skutkuje uzyskaniem nieskończonych ocen parametrów modelu [Albert i Anderson 1984, s. 3–6; Firth 1993, s. 31–32;

Heinze 1999, s. 4–5]. Szansa pojawienia się całkowitego rozdzielenia jest tym większa, im mniejszy zbiór danych jest poddawany analizie.

Pierwszą metodą badaną w pracy [Fijorek 2012] jest propozycja G. Heinzego i M. Schempera [2002] oparta na wynikach badań D. Firtha [1993]. Drugą propozycją jest metoda HLM (*hidden logistic model*) przedstawiona przez P.J. Rousseeuwa i A. Christmanna [2003] (w okresie późniejszym G. Tutz i F. Leitenstorfer [2006] opublikowali pracę o podobnym charakterze). Wnioski płynące z artykułu [Fijorek 2012] można podsumować w następujący sposób: oceny parametrów uzyskane metodą Firtha oraz HLM charakteryzują się mniejszym obciążeniem w porównaniu z ocenami MNW (w zakresie przeprowadzonych symulacji), przy czym metoda Firtha znacznie lepiej redukuje obciążenie niż HLM. Ponadto w modelu Firtha przedziały ufności wyznaczane za pomocą metody *profile likelihood* (PL) w porównaniu z przedziałami Walda uzyskują bliższy nominalnemu poziom pokrycia.

Praktyczna użyteczność modelu Firtha uzasadnia dalsze badania jego właściwości. Celem artykułu jest zaproponowanie metod aproksymacji modelu Firtha oraz wskazanie obszarów ich zastosowań w analizie danych empirycznych.

2. Model regresji logistycznej Firtha

W standardowym modelu regresji logistycznej przyjmuje się, że zmienna zależna $y_i \in \{0,1\}\,(i=1,\ldots,n)$ podlega rozkładowi Bernoulliego z prawdopodobieństwem sukcesu $F(\mathbf{x}_i'\boldsymbol{\theta})$, gdzie $F(\cdot)$ jest dystrybuantą rozkładu logistycznego, x_i to p-wymiarowy wektor zmiennych objaśniających, a $\boldsymbol{\theta} \in \mathbb{R}^p$ to (zawierający wyraz wolny) p-wymiarowy wektor parametrów strukturalnych [Long 1997, s. 40–61].

W celu oszacowania parametrów modelu wyznacza się funkcję wiarygodności oraz jej logarytm:

$$L(\boldsymbol{\theta} \mid y_1,\ldots,y_n) = \prod_{i=1}^{n} F(\mathbf{x}_i'\boldsymbol{\theta})^{y_i}\left[1-F(\mathbf{x}_i'\boldsymbol{\theta})\right]^{1-y_i},$$

$$l(\boldsymbol{\theta} \mid y_1,\ldots,y_n) = \sum_{i=1}^{n} y_i \ln F(\mathbf{x}_i'\boldsymbol{\theta}) + (1-y_i)\ln\left[1-F(\mathbf{x}_i'\boldsymbol{\theta})\right].$$

Następnie oblicza się pochodne cząstkowe logarytmu funkcji wiarygodności względem parametrów modelu:

$$s(\boldsymbol{\theta} \mid y_1,\ldots,y_n) = \sum_{i=1}^{n} \left(y_i - F(\mathbf{x}_i'\boldsymbol{\theta})\right)\mathbf{x}_i.$$

Rozwiązanie układu równań $s(\boldsymbol{\theta} \mid y_1,\ldots,y_n) = 0$ jest równoważne ze znalezieniem ocen parametrów maksymalizujących funkcję wiarygodności.

W przypadku modelu regresji logistycznej Firtha $s(\boldsymbol{\theta} \mid y_i, \ldots, y_n)$ funkcję zastępuje się funkcją:

$$s^*(\boldsymbol{\theta} \mid y_1, \ldots, y_n) = \sum_{i=1}^{n} \left(y_i - F(\mathbf{x}'_i\boldsymbol{\theta}) + h_i \left(\tfrac{1}{2} - F(\mathbf{x}'_i\boldsymbol{\theta}) \right) \right) \mathbf{x}_i,$$

gdzie h_i to diagonalne elementy macierzy $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$, $\mathbf{X}$ to macierz danych, a $\mathbf{W}$ jest macierzą diagonalną o wymiarach $n \times n$, której i-ty diagonalny element jest równy $F(\mathbf{x}'_i\boldsymbol{\theta})(1 - F(\mathbf{x}'_i\boldsymbol{\theta}))$. Modyfikacja postaci funkcji $s(\boldsymbol{\theta} \mid y_1, \ldots, y_n)$ jest tożsama z modyfikacją funkcji wiarygodności:

$$L^*(\boldsymbol{\theta} \mid y_1, \ldots, y_n) = L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) \left| \mathbf{I}_{\boldsymbol{\theta}} \right|^{\frac{1}{2}},$$

gdzie $\mathbf{I}_{\boldsymbol{\theta}}$ to macierz informacyjna postaci [Greene 2003, s. 670–673]:

$$\mathbf{I}_{\boldsymbol{\theta}} = -\sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{i=1}^{n} F(\mathbf{x}'_i\boldsymbol{\theta})(1 - F(\mathbf{x}'_i\boldsymbol{\theta}))\mathbf{x}_i\mathbf{x}'_i.$$

$L^*(\boldsymbol{\theta} \mid y_1, \ldots, y_n)$ bywa nazywana funkcją wiarygodności z karą (*penalized likelihood function*) [Heinze i Schemper 2002, Heinze 2006]. D. Firth [1993] udowodnił, że opisane powyżej podejście, poza rozwiązaniem problemu całkowitego rozdzielenia, usuwa również obciążenie pierwszego rzędu towarzyszące metodzie największej wiarygodności.

3. Model regresji logistycznej z ukrytą zmienną objaśnianą (HLM)

W celu rozwiązania problemu całkowitego rozdzielenia P.J. Rousseeuw i A. Christmann [2003] proponują natomiast założyć istnienie mechanizmu stochastycznego, który powoduje, że prawdziwe wartości zmiennej zależnej (y_i) są nieobserwowalne. W tym ujęciu przyjmuje się, że sukces jest obserwowany z prawdopodobieństwem δ_1, co przekłada się na możliwość błędnego uznania sukcesu za porażkę z prawdopodobieństwem $1 - \delta_1$. Podobnie porażka jest obserwowana z prawdopodobieństwem $1 - \delta_0$, a jej błędna klasyfikacja występuje z prawdopodobieństwem δ_0.

P.J. Rousseeuw i A. Christmann [2003, s. 317–318] pokazują, że estymacja parametrów modelu HLM sprowadza się do zastąpienia oryginalnych obserwacji pseudoobserwacjami $\tilde{y}_i$ obliczonymi w następujący sposób:

$$\tilde{y}_i = (1 - y_i)\delta_0 + y_i\delta_1.$$

Następnie do tak utworzonych pseudoobserwacji można zastosować klasyczną metodę największej wiarygodności, której celem jest maksymalizacja funkcji wiarygodności względem wektora parametrów $\boldsymbol{\theta} \in \mathbb{R}^p$:

$$L(\boldsymbol{\theta} \mid \tilde{y}_1, \ldots, \tilde{y}_n) = \prod_{i=1}^{n} F(\mathbf{x}'_i\boldsymbol{\theta})^{\tilde{y}_i} \left[1 - F(\mathbf{x}'_i\boldsymbol{\theta}) \right]^{1-\tilde{y}_i}.$$

Jak podają P.J. Rousseeuw i A. Christmann [2003, s. 320–323], precyzyjne wartości δ_0 oraz δ_1 mogą być poznane jedynie dla bardzo dużych zbiorów danych. W przypadku małych zbiorów danych można przyjąć, że $\delta_0 = 0{,}01$ i $\delta_1 = 0{,}99$. Jeśli założymy, że suma pseudoobserwacji powinna być równa sumie oryginalnych wartości zmiennej objaśnianej:

$$\hat{\pi} = \tfrac{1}{n}\sum_{i=1}^{n} \tilde{y}_i = \tfrac{1}{n}\sum_{i=1}^{n} y_i,$$

δ_0 oraz δ_1 przyjmują wartości:

$$\delta_0 = \frac{\tilde{\pi}\delta}{1+\delta} \quad \text{oraz} \quad \delta_1 = \frac{1+\tilde{\pi}\delta}{1+\delta}.$$

P.J. Rousseeuw i A. Christmann [2003, s. 322] zakładają, że $\delta = 0{,}01$ (w dalszej części pracy rezultaty oznaczone skrótem HLM wynikają z tego założenia). Należy zaznaczyć, że opisane sposoby doboru δ_0 oraz δ_1 nie mają silnego uzasadnienia teoretycznego. Można sądzić, że algorytm określania optymalnych wartości δ_0 oraz δ_1 powinien uwzględniać rozmiar próby oraz liczbę zmiennych objaśniających. Zaprezentowane w dalszej części artykułu wyniki badań dostarczają nowych informacji w tym zakresie.

4. Alternatywne spojrzenie na model Firtha oraz HLM

4.1. Propozycje aproksymacji modelu Firtha

Jak już wspomniano wcześniej, model regresji logistycznej Firtha może być zdefiniowany jako standardowy model regresji logistycznej ze zmodyfikowaną funkcją wiarygodności. Na model Firtha można też jednak spojrzeć z innej perspektywy i wykazać, że podejście to sprowadza się do zastąpienia każdej oryginalnej obserwacji y_i dwiema nowymi obserwacjami, tj. y_i oraz $1 - y_i$ z wagami wynoszącymi odpowiednio $1 + \frac{h_i}{2}$ oraz $\frac{h_i}{2}$ [Heinze i Schemper 2002, s. 2412]. W rezultacie otrzymuje się zbiór danych liczący $2n$ przypadków, w którym:

$$\mathbf{y}^* = (y_1, \ldots, y_n, 1 - y_1, \ldots, 1 - y_n),$$

$$\mathbf{X}^{*\prime} = (\mathbf{x}'_1, \ldots, \mathbf{x}'_n, \mathbf{x}'_1, \ldots, \mathbf{x}'_n),$$

a wagi poszczególnych przypadków to $\mathbf{h} = \left(1 + \frac{h_1}{2}, \ldots, 1 + \frac{h_n}{2}, \frac{h_1}{2}, \ldots, \frac{h_n}{2}\right)$.

W kontekście modelu Firtha powyższa obserwacja nie ułatwia obliczeniowego aspektu procesu estymacji, gdyż nadal wagi są zależne od danych oraz –

co bardziej istotne – wagi zmieniają się wraz z kolejnymi iteracjami algorytmu optymalizującego funkcję wiarygodności. Korzystając z powyższego spostrzeżenia, można jednak zaproponować metodę aproksymacji modelu Firtha polegającą na wyznaczeniu stałych wag, które aproksymują element $\frac{h_i}{2}$. Jak wykazują D.W. Hosmer i S. Lemeshow [2000, s. 169]:

$$\sum_{i=1}^{n} h_i = p,$$

potencjalnym sposobem przybliżania h_i jest zatem uznanie, że $h_i \approx \frac{p}{n} = \bar{h}$.

Pierwsza propozycja aproksymacji modelu Firtha (aFirth) polega na zastosowaniu standardowej metody największej wiarygodności do zbioru danych $(\mathbf{y}^*, \mathbf{X}^*)$, w którym wagi pierwszych n obserwacji to $1 + \frac{\bar{h}}{2}$, a wagi kolejnych n obserwacji to $\frac{\bar{h}}{2}$. Konstrukcja wag powoduje, że rola sztucznych obserwacji maleje wraz ze wzrostem n, a przy założeniu stałego rozmiaru zbioru danych ich rola rośnie wraz ze wzrostem liczby zmiennych niezależnych. Ponadto należy podkreślić, że oryginalne obserwacje odgrywają dominującą rolę w procesie estymacji, a obserwacje sztuczne stosunkowo niewielką, np. jeśli $n = 100$ oraz $p = 5$, to waga sztucznej obserwacji wynosi 0,025, natomiast obserwacji oryginalnej 1,025.

Jak wynika z obliczeń autora (inspirowanych pracą [Tutz i Leitenstorfer 2006]), również model HLM, który w swej podstawowej formie zakłada transformację oryginalnych obserwacji do pseudoobserwacji, można zmodyfikować i zapisać w formie zbioru danych $(\mathbf{y}^*, \mathbf{X}^*)$ z wagami w_i:

$$\begin{cases} y_i = 1 \rightarrow w_i = \delta_1 \\ y_i = 0 \rightarrow w_i = 1 - \delta_0 \end{cases} \quad \text{dla} \quad i = 1, \ldots, n,$$

$$\begin{cases} 1 - y_i = 1 \rightarrow w_i = \delta_0 \\ 1 - y_i = 0 \rightarrow w_i = 1 - \delta_1 \end{cases} \quad \text{dla} \quad i = n+1, \ldots, 2n.$$

Druga propozycja aproksymacji modelu Firtha polega na założeniu, że $\delta = \frac{\bar{h}}{2}$, oraz zastosowaniu standardowej metody największej wiarygodności do zbioru danych $(\mathbf{y}^*, \mathbf{X}^*)$ z wagami w_i. Propozycja ta będzie oznaczana jako zmodyfikowany HLM (mHLM).

4.2. Aspekt praktyczny aproksymacji modelu Firtha

Aby uzyskać oceny parametrów modelu regresji logistycznej za pomocą metody Firtha, trzeba posiadać specjalne oprogramowanie. Przykładem implementacji modelu Firtha jest biblioteka *logistf* [Heinze i Ploner 2004] dostępna w środowisku obliczeń statystycznych R [*R Development Core Team…* 2010]. Inne

pakiety statystyczne (z pewnymi wyjątkami) nie udostępniają modelu Firtha, lecz większość z nich dopuszcza ważenie obserwacji niecałkowitymi wagami. W tej sytuacji użytkownik może aproksymować model Firtha, stosując standardowy model regresji logistycznej (w środowisku R służy temu funkcja *glm*) z wagami obliczonymi na podstawie wyników niniejszego artykułu.

Istniejące implementacje modelu Firtha znajdują się we wczesnym stadium rozwoju i w rezultacie charakteryzują się stosunkowo ograniczoną funkcjonalnością. Przykładem może być wspomniana powyżej biblioteka *logistf*, która umożliwia estymację modelu oraz testowanie hipotez statystycznych, lecz już automatyczne metody doboru zmiennych nie są obsługiwane. Proponowana aproksymacja umożliwia podejście dwuetapowe, tzn. w pierwszym kroku użytkownik stosuje aproksymację modelu Firtha, uzyskując tym samym dostęp do funkcjonalności standardowych narzędzi estymacji regresji logistycznej, natomiast w drugim etapie użytkownik dokonuje estymacji finalnego modelu za pomocą procedury Firtha.

5. Badania symulacyjne oraz przykład empiryczny

5.1. Założenia badań symulacyjnych

W tej części opracowania za pomocą badań symulacyjnych następujące trzy metody estymacji modelu regresji logistycznej:

– metoda aproksymowanego Firtha (aFirth),
– zmodyfikowana metoda HLM (mHLM),
– standardowa metoda HLM,

zostaną porównane z wzorcem, którym w tym badaniu jest model Firtha. W porównaniach wykorzystano zbiory danych zawierające:

– wariant A – k ($k = 2, 4, 6, 8, 10, 12$) nieskorelowanych zmiennych losowanych z rozkładu N (0, 1),
– wariant B – $\frac{k}{2}$ nieskorelowanych zmiennych losowanych z rozkładu N (0, 1) oraz $\frac{k}{2}$ zmiennych zero-jedynkowych losowanych z rozkładu Bernoulliego (prawdopodobieństwo uzyskania 1 przyjęto na poziomie 0,5).

W toku symulacji generowano R = 1000 zestawów danych liczących 50, 75, 100 oraz 200 przypadków. Wartość y_i określano jako losową realizację z rozkładu Bernoulliego, w którym prawdopodobieństwo sukcesu jest równe $F(\mathbf{x}'_i\boldsymbol{\theta})$. W ramach badania symulacyjnego wektor parametrów modelu rozważono w dwóch wariantach:

– wariant 1 – $\theta_j = 1 \quad (j = 0, \ldots, k)$,
– wariant 2 – $\theta_j = 0{,}5 \quad (j = 0, \ldots, k)$,

gdzie θ_0 oznacza wyraz wolny. W rezultacie otrzymano cztery warianty symulacji oznaczone kolejno A1, A2, B1, B2.

Zachowanie się ocen parametrów określano, mierząc ich przeciętne względne obciążenie zgodnie ze wzorem:

$$B = \frac{1}{k}\sum_{j=0}^{k}\frac{\left[\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_j^r - \theta_j\right]}{\theta_j},$$

gdzie $\hat{\theta}_j^r$ to ocena *j*-tego parametru na podstawie *r*-tej symulacji.

5.2. Wyniki badań symulacyjnych

Wyniki przeprowadzonych badań symulacyjnych zestawiono w tabelach 1–4. Zgodnie z oczekiwaniami zaobserwowano, że ogólnie im mniejszy jest zbiór danych przy założeniu stałej liczby zmiennych objaśniających oraz im więcej zmiennych objaśniających przy założeniu stałego rozmiaru zbioru danych, tym większe jest obciążenie ocen parametrów. Ponadto – również zgodnie z oczekiwaniami – oceny parametrów modelu regresji logistycznej wyznaczane metodą Firtha charakteryzowały się bardzo niskim obciążeniem w przekroju niemal wszystkich wariantów symulacji. Niewielkie odstępstwa od braku obciążenia zaobserwowano w najbardziej ekstremalnych przypadkach, tzn. w przypadku $n = 50$, $n = 75$ oraz $k = 10$, $k = 12$. Wykonane symulacje pozwalają przewidywać, że korekta ocen parametrów modelu regresji logistycznej o obciążenie rzędu wyższego niż pierwszy nie będzie skutkować poprawą, którą można by uznać za istotną z praktycznego punktu widzenia. Interesujące rozważania teoretyczne na ten temat w klasie uogólnionych modeli liniowych (*generalized linear model*), których szczególnym przypadkiem jest model regresji logistycznej, przedstawili G. Cordeiro i L. Barroso [2007].

Model HLM w swej standardowej postaci w porównaniu z modelem Firtha wykazał bardzo słabą kontrolę obciążenia ocen parametrów. Dopiero przy $n = 200$ oraz niewielkiej liczbie zmiennych objaśniających obciążenie znajdowało się na akceptowalnym poziomie. Dużo lepszą kontrolą obciążenia wykazał się natomiast zmodyfikowany model HLM (mHLM), który w znacznej części symulacji dawał wyniki zbliżone do modelu Firtha. Interesujące jest, że propozycja aFirth nie dostarczyła zadowalających rezultatów. Aproksymacja ta, ogólnie rzecz biorąc, prowadziła do zbyt mocnej korekty obciążenia (oceny parametrów były mniejsze, niż powinny być), przy czym zjawisko było tym bardziej dotkliwe, im więcej zmiennych objaśniających znajdowało się w modelu (przy założeniu stałego rozmiaru zbioru danych).

Tabela 1. Wartości przeciętnego względnego obciążenia ocen parametrów – wariant A1 (w %)

Liczba obserwacji	Model	Liczba zmiennych (razem z wyrazem wolnym)					
		3	5	7	9	11	13
50	HLM	12,2	23,6	41,1	83,5	141,7	187,7
	mHLM	7,9	7,1	1,4	–4,6	–11,8	–21,1
	Firth	1,3	2,1	1,5	1,5	–0,4	–11,8
	aFirth	–2,6	–11,5	–22,6	–32,2	–40,2	–48,0
75	HLM	6,6	11,6	18,2	29,1	49,6	73,2
	mHLM	4,8	4,4	1,6	–2,5	–6,5	–13,0
	Firth	0,5	0,6	0,4	1,3	3,8	2,3
	aFirth	–1,6	–7,7	–15,7	–23,8	–30,7	–38,0
100	HLM	3,3	7,3	10,6	18,1	26,8	37,5
	mHLM	2,5	3,1	1,0	–0,3	–4,5	–8,8
	Firth	–0,7	0,1	–0,3	1,5	2,7	3,5
	aFirth	–2,1	–6,0	–12,2	–18,0	–25,3	–31,4
200	HLM	1,5	2,5	4,0	4,9	7,4	9,1
	mHLM	1,9	1,9	1,4	–0,2	–1,5	–3,9
	Firth	0,3	0,4	0,6	0,2	0,8	0,4
	aFirth	–0,4	–2,6	–5,5	–9,8	–13,8	–18,5

Źródło: opracowanie własne.

Tabela 2. Wartości przeciętnego względnego obciążenia ocen parametrów – wariant A2 (w %)

Liczba obserwacji	Model	Liczba zmiennych (razem z wyrazem wolnym)					
		3	5	7	9	11	13
50	HLM	6,2	14,5	24,7	39,7	66,8	118,5
	mHLM	4,0	7,0	8,8	9,5	7,8	5,6
	Firth	–2,3	–0,6	0,8	2,9	3,9	6,0
	aFirth	–2,1	–4,2	–7,5	–12,1	–18,3	–23,8
75	HLM	5,5	9,9	12,7	19,4	25,9	37,6
	mHLM	4,4	6,2	5,2	6,3	5,6	5,4
	Firth	0,2	0,9	–0,2	1,0	1,2	3,0
	aFirth	0,5	–0,9	–5,0	–7,8	–11,8	–15,7
100	HLM	2,9	6,8	9,0	12,5	16,5	21,2
	mHLM	2,4	4,6	4,4	4,5	4,3	3,6
	Firth	–0,7	0,6	0,2	0,4	0,6	0,8
	aFirth	–0,4	–0,5	–3,1	–5,8	–8,9	–12,3
200	HLM	0,6	2,8	4,0	4,9	6,4	6,9
	mHLM	0,8	2,5	2,8	2,5	2,6	1,4
	Firth	–0,7	0,5	0,6	0,3	0,5	–0,4
	aFirth	–0,5	0,1	–0,9	–2,5	–3,9	–6,5

Źródło: opracowanie własne.

Tabela 3. Wartości przeciętnego względnego obciążenia ocen parametrów – wariant B1 (w %)

Liczba obserwacji	Model	Liczba zmiennych (razem z wyrazem wolnym)					
		3	5	7	9	11	13
50	HLM	13,4	31,6	87,5	186,8	274,1	354,0
	mHLM	9,9	14,8	20,0	17,4	10,7	1,1
	Firth	0,3	0,7	1,6	–6,4	–19,6	–34,4
	aFirth	–4,0	–15,8	–29,2	–41,8	–52,0	–59,9
75	HLM	7,8	16,1	36,0	85,1	173,5	263,2
	mHLM	6,4	10,1	15,6	20,6	23,4	21,0
	Firth	0,7	0,8	2,1	1,9	–3,4	–16,4
	aFirth	–2,1	–10,5	–21,7	–33,5	–44,6	–53,8
100	HLM	5,2	10,3	18,6	38,7	86,2	181,1
	mHLM	4,6	7,2	10,1	14,7	20,3	25,8
	Firth	0,4	0,7	0,3	1,0	0,5	–2,7
	aFirth	–1,6	–7,7	–17,2	–28,0	–38,6	–47,7
200	HLM	1,4	3,3	7,0	10,1	16,6	27,2
	mHLM	1,7	2,8	5,3	6,2	9,0	12,4
	Firth	–0,3	–0,2	0,8	0,0	0,3	0,5
	aFirth	–1,2	–4,2	–8,7	–16,2	–24,4	–33,0

Źródło: opracowanie własne.

Tabela 4. Wartości przeciętnego względnego obciążenia ocen parametrów – wariant B2 (w %)

Liczba obserwacji	Model	Liczba zmiennych (razem z wyrazem wolnym)					
		3	5	7	9	11	13
50	HLM	9,8	18,0	29,1	71,1	163,6	232,0
	mHLM	7,9	11,9	14,8	22,3	28,2	29,4
	Firth	1,0	1,7	0,5	3,8	4,3	–2,1
	aFirth	0,7	–3,3	–10,8	–17,8	–26,5	–35,1
75	HLM	4,2	11,0	14,9	25,0	41,5	102,8
	mHLM	3,4	8,2	9,1	13,8	17,1	27,1
	Firth	–0,9	1,6	0,2	1,6	1,1	4,5
	aFirth	–1,0	–1,3	–6,7	–11,3	–18,9	–25,4
100	HLM	4,3	6,6	10,4	16,7	20,8	35,9
	mHLM	4,0	5,0	7,0	10,4	10,5	16,0
	Firth	0,7	0,2	0,4	1,7	–0,5	0,9
	aFirth	0,6	–1,7	–4,5	–8,0	–15,1	–20,9
200	HLM	1,5	2,6	3,5	6,1	8,9	10,6
	mHLM	1,7	2,3	2,6	4,4	6,1	6,6
	Firth	0,1	–0,1	–0,5	0,2	0,8	–0,3
	aFirth	0,1	–0,8	–2,7	–4,2	–6,6	–11,3

Źródło: opracowanie własne.

Podsumowując wyniki symulacji, należy stwierdzić, że model Firtha powinien być zawsze przedkładany nad inne rozważane w artykule podejścia. W sytuacji gdy jego zastosowanie nie jest możliwe, należy rekomendować stosowanie zmodyfikowanego modelu HLM (mHLM). W przypadku niewielkiej liczby zmiennych niezależnych różnica pomiędzy dwoma podejściami nie powinna być duża. Dla zbiorów danych liczących 200 i więcej obserwacji różnice pomiędzy podejściami przestają być zauważalne przy założeniu zrównoważonego stosunku liczby sukcesów do porażek.

5.3. Przykład empiryczny

W tabeli 5 zestawiono wyniki estymacji modelu regresji logistycznej za pomocą omówionych w artykule metod. Zbiór danych, który posłużył do estymacji modeli omawiają i analizują P.J. Rousseeuw i A. Christmann [2003, s. 327–328], a jego oryginalnym źródłem jest praca [Finney 1947]. Zmienną objaśnianą jest zmienna symbolizująca wystąpienie u badanej osoby wazokonstrykcji (termin oznacza zwężenie światła naczyń krwionośnych). Zmiennymi objaśniającymi są objętość powietrza wdychanego (*volume*) oraz tempo wdychania powietrza (*rate*).

Tabela 5. Wyniki estymacji modelu regresji logistycznej dla zbioru danych

Zmienna	MNW		HLM		mHLM		Firth		aFirth	
	$\hat{\boldsymbol{\theta}}$	$SE(\hat{\boldsymbol{\theta}})$	$\hat{\boldsymbol{\theta}}$	$SE(\hat{\boldsymbol{\theta}})$	$\hat{\boldsymbol{\theta}}$	$SE(\hat{\boldsymbol{\theta}})$	$\hat{\boldsymbol{\theta}}$	$SE(\hat{\boldsymbol{\theta}})$	$\hat{\boldsymbol{\theta}}$	$SE(\hat{\boldsymbol{\theta}})$
Wyraz wolny	–9,53	3,23	–9,13	3,09	–8,21	2,78	–7,83	2,66	–7,27	2,38
Volume	2,88	1,49	3,71	1,37	3,31	1,23	3,17	1,18	2,90	1,06
Rate	2,65	0,91	2,55	0,88	2,31	0,80	2,18	0,76	2,06	0,70

$SE(\hat{\boldsymbol{\theta}})$ – błąd średni szacunku.

Źródło: opracowanie własne na podstawie: [Finney 1947].

Analizując wyniki estymacji, można stwierdzić, że do ocen metody Firtha najbardziej podobne są oceny zmodyfikowanej metody HLM, następnie aproksymowanej metody Firtha, dalej standardowej metody HLM, a najgorzej w tym porównaniu wypadają oceny metody największej wiarygodności.

6. Podsumowanie

W artykule przedstawiono model regresji logistycznej Firtha w kontekście wag przypisywanych przez metodę poszczególnym obserwacjom ze zbioru

danych oraz zaprezentowano przekształcenie modelu HLM do podobnej postaci. Na podstawie wniosków płynących z alternatywnego spojrzenia na model Firtha oraz HLM zaproponowano dwie metody aproksymacji modelu Firtha. Następnie symulacyjnie zbadano jakość aproksymacji oraz omówiono praktyczne korzyści płynące z jej stosowania.

Literatura

Albert A., Anderson J.A. [1984], *On the Existence of Maximum Likelihood Estimates in Logistic Regression Models*, „Biometrika", vol. 71.

Cordeiro G., Barroso L. [2007], *A Third-order Bias Corrected Estimate in Generalized Linear Models*, „Test", vol. 16, nr 1.

Fijorek K. [2012], *Porównanie modeli regresji logistycznej odpornych na problem całkowitego rozdzielenia*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, nr 884, Kraków.

Finney D.J. [1947], *The Estimation from Individual Records of the Relationship between Dose and Quantal Response*, „Biometrika", vol. 34.

Firth D. [1993], *Bias Reduction of Maximum Likelihood Estimates*, „Biometrika", vol. 80.

Greene W.H. [2003], *Econometric Analysis*, Pearson Education, New Jersey.

Heinze G. [1999], *The Application of Firth's Procedure to Cox and Logistic Regression*, Technical Report 10, Department of Medical Computer Sciences, Section of Clinical Biometrics, Vienna University, Vienna.

Heinze G. [2006], *A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data*, „Statistics in Medicine", vol. 25.

Heinze G., Ploner M. [2004], *A SAS Macro, S-PLUS Library and R Package to Perform Logistic Regression without Convergence Problems*, Technical Report 2, Section of Clinical Biometrics, Department of Medical Computer Sciences, Medical University of Vienna, Vienna.

Heinze G., Schemper M. [2002], *A Solution to the Problem of Separation in Logistic Regression*, „Statistics in Medicine", vol. 21.

Hosmer D.W., Lemeshow S. [2000], *Applied Logistic Regression*, John Wiley and Sons.

Long J.S. [1997], *Regression Models for Categorical and Limited Dependent Variables*, Sage, Thousand Oaks.

R Development Core Team. R: A Language and Environment for Statistical Computing [2010], R Foundation for Statistical Computing, Vienna.

Rousseeuw P.J., Christmann A. [2003], *Robustness against Separation and Outliers in Logistic Regression*, „Computational Statistics and Data Analysis", vol. 43.

Tutz G., Leitenstorfer F. [2006], *Response Shrinkage Estimators in Binary Regression*, „Computational Statistics and Data Analysis", vol. 50.

Firth's Logistic Regression Approximation by Weighting Observations

Firth's approach to a logistic regression is presented from the perspective of weighted data points. Hidden Logistic Model is reformulated accordingly and two approximations of Firth's procedure are introduced. A simulation study was conducted to investigate and compare the quality of the approximations.

Keywords: logistic regression, bias reduction, complete separation, approximate estimation.