PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Beata Jackowska, Ewa Wycinka
Uniwersytet Gdański

MODELE RYZYKA SKREŚLENIA Z LISTY STUDENTÓW NA PRZYKŁADZIE STUDENTÓW TRYBU NIESTACJONARNEGO

1. Wstęp

Każdego roku część studentów przerywa studia na różnym etapie ich zaawansowania. Zmniejszająca się liczba studentów na kolejnych semestrach ma niekorzystny wpływ na planowanie i efektywne wykorzystanie kadry i środków technicznych uczelni, a także na możliwość prognozowania strumieni finansowych. Z tego powodu dla władz uczelni ważne jest rozpoznanie i pomiar czynników ryzyka przerwania studiów przez studenta.

Na przykładzie rozkładu odejść ze studiów studentów z kohorty naboru na rok akademicki 2004/2005 trybu zaocznego Wydziału Zarządzania Uniwersytetu Gdańskiego (UG) podjęta została próba wyodrębnienia czynników wpływających na ryzyko przerwania studiów. Celem tego opracowania jest skonstruowanie modelu, który mógłby służyć prognozowaniu natężenia i rozkładu odejść w kolejnych kohortach studentów.

2. Dobór zmiennych objaśniających

Skreślenie z listy studentów jest dychotomiczną zmienną objaśnianą (zdarzenie wystąpiło lub nie wystąpiło) reprezentowaną przez zmienną o rozkładzie zero-jedynkowym. Za potencjalne zmienne objaśniające przyjęto: płeć, wiek, odległość miejsca zamieszkania od uczelni, kierunek studiów, wcześniejsze studiowanie innego kierunku na UG[1]. Ciągłe zmienne objaśniające zostały poddane próbie kategoryzacji. Analiza wykresów ilorazów szans [Williams i in. 2006] wskazała na istotne zmiany w przebiegu krzywych ilorazów szans w dwóch przypadkach: dla kobiet studiujących na semestrze pierwszym odległość została podzielona na cztery kategorie według kwartyli, natomiast zmienna „wiek" dla mężczyzn z semestru pierwszego

[1] Liczba lat, które upłynęły od zdania matury, okazała się zmienną niepowiązaną ze zmienną objaśnianą.

została zdychotomizowana według trzeciego kwartyla[2]. Prawdopodobieństwa testów niezależności chi-kwadrat dla wybranych w przedstawiony sposób punktów odcięcia (*cutpoint*) wykazały najwyższą efektywność tego podziału [Williams i in. 2006]. Następnie za pomocą analizy log-liniowej zbadano wpływ wszystkich skategoryzowanych zmiennych objaśniających na zmienną objaśnianą. Powyższe czynności zostały przeprowadzone w odniesieniu do wszystkich semestrów w ocenie modeli łącznych oraz w rozbiciu na osobne modele dla obu płci. Za kryterium doboru zestawu zmiennych objaśniających przyjęto kryterium informacyjne Akaikego dla modeli log-liniowych AIC (*Akaike information criterion*) [Agresti 2002].

3. Modele prawdopodobieństwa zdarzenia

Zmienną objaśnianą jest zmienna dychotomiczna Y o wartościach:

$$Y = \begin{cases} 1 & \text{zdarzenie wystąpiło} \\ 0 & \text{zdarzenie nie wystąpiło} \end{cases}.$$

Modelowaniu podlega prawdopodobieństwo warunkowe wystąpienia interesującego nas zdarzenia (sukcesu), pod warunkiem że zmienne niezależne przyjęły wartości $x_1, x_2, \ldots, x_n$. Przyjmuje się, że prawdopodobieństwo to jest funkcją liniowej kombinacji wartości zmiennych objaśniających $z = b_0 + \sum_{i=1}^{n} b_i x_i$:

$$p = P(Y = 1 \mid x_1, x_2, \ldots, x_n) = F(z) = F\left(b_0 + \sum_{i=1}^{n} b_i x_i\right).$$

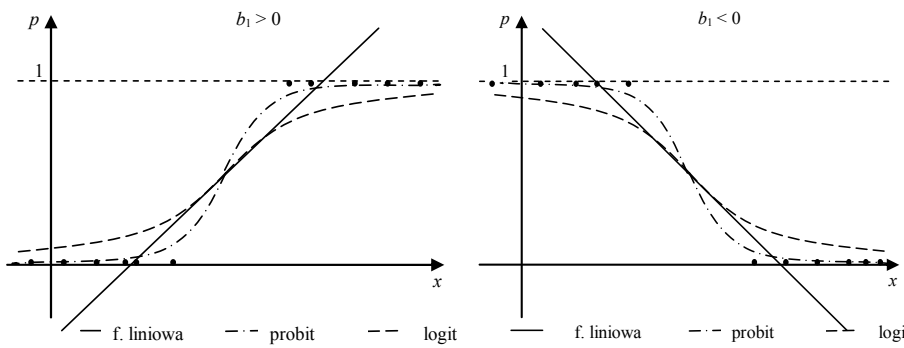
Aby zagwarantować spełnienie warunku $0 \le p \le 1$, za funkcję F można przyjąć dystrybuantę pewnego rozkładu. Najczęściej stosowane są modele:

- Liniowy ucięty model prawdopodobieństwa[3]: $p = F(z) = \begin{cases} 0 & \text{dla} \quad z \le 0 \\ z & \text{dla} \quad 0 < z < 1 \\ 1 & \text{dla} \quad z \ge 1 \end{cases}$.

[2] Optymalnym punktem odcięcia zmiennej „wiek" dla kobiet był dziewięćdziesiąty piąty percentyl, jednak ostatecznie model logitowy uwzględniający zmienną „wiek" jako zmienną ciągłą okazał się lepszy według kryterium Akaikego (AIC = 128,845) w porównaniu z modelem ze zmienną dychotomiczną „wiek" (AIC = 131,912). Analogiczna sytuacja wystąpiła w przypadku modelu probitowego (zob. tab. 1).

[3] Parametry modelu interpretuje się tak, jak parametry w liniowym modelu regresji. Jednakże interpretacja ma sens jedynie wówczas, gdy ocena parametru mieści się w przedziale $[-1, 1]$. Przy szacowaniu parametrów modelu, ze względu na heteroskedastyczność składnika losowego, stosuje się ważoną metodę najmniejszych kwadratów. Wraz z rozwojem technik obliczeniowych model liniowy stracił swoje znaczenie i obecnie w literaturze traktowany jest jako punkt wyjścia do opisu modeli bardziej użytecznych w praktyce [Agresti 2002; Gruszczyński 2001; Harrell 2001].

- Model logitowy[4]: $p = F(z) = \dfrac{1}{1+e^{-z}} = \dfrac{e^{z}}{1+e^{z}}$.
- Model probitowy[5]: $p = F(z) = \int_{-\infty}^{z} \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{t^2}{2}\right) dt$.



Rys. 1. Porównanie modeli prawdopodobieństwa $p = F(b_0 + b_1 x)$ w przypadku jednej zmiennej objaśniającej

Źródło: opracowanie własne.

Przebieg powyższych funkcji dla jednej zmiennej objaśniającej przedstawia wykres na rys. 1.

4. Miary dopasowania modelu prawdopodobieństwa

Przedstawione modele mogą służyć dwóm celom:

- oszacowaniu prawdopodobieństwa $\hat{p}$ pewnego zdarzenia,
- oszacowaniu objaśnianej zmiennej zero-jedynkowej $\hat{Y}$.

W pierwszym przypadku do oceny modelu wykorzystuje się miary dopasowania typu R^2 (tzw. pseudo-R^2), m.in.: kwadrat współczynnika korelacji między wartościami obserwowanymi y_i a oszacowanymi wartościami prawdopodobieństwa $\hat{p}_i$, R^2

[4] Do interpretacji parametrów logistycznej funkcji regresji wykorzystuje się pojęcie ilorazu szans (szansę określa się jako stosunek prawdopodobieństwa wystąpienia zdarzenia do prawdopodobieństwa niewystąpienia zdarzenia). Jeżeli wartość zmiennej objaśniającej x_i wzrośnie o jednostkę (przy stałych wartościach pozostałych zmiennych), to iloraz szans zmieni się $\exp(b_i)$-krotnie (wzrośnie, gdy $\exp(b_i) > 1$; zmaleje, gdy $\exp(b_i) < 1$). Estymatory b_i parametrów funkcji regresji wyznacza się metodą największej wiarygodności.

[5] Interpretacja parametrów modelu probitowego sprowadza się jedynie do stwierdzenia, czy dana zmienna jest stymulantą (gdy $b_i > 0$), czy destymulantą modelu (gdy $b_i < 0$). Oszacowanie parametrów b_i otrzymuje się metodą największej wiarygodności.

Efrona = $1-\frac{N}{n_0 \cdot n_1}\sum\left(y_i-\hat{p}_i\right)^2$, gdzie N jest liczbą obserwacji, wśród których zanotowano n_0 zer i n_1 jedynek. Dla modeli logitowych i probitowych szczególne znaczenie mają miary pseudo-R^2 oparte na wartościach funkcji wiarygodności:

$$R^2\text{McFaddena}=1-\ln L_n/\ln L_0\ ,$$

$$R^2\text{ Cragga-Uhlera} = \left(1-\left(L_0/L_n\right)^{2/n}\right)\Big/\left(1-\left(L_0\right)^{2/n}\right),$$

$$R^2\text{ Vealla-Zimmermanna}^6 = \frac{2\left(\ln L_n-\ln L_0\right)}{2\left(\ln L_n-\ln L_0\right)+N}\cdot\frac{N-2\ln L_0}{-2\ln L_0},$$

$$R^2\text{ Estrelli}=1-\left(\ln L_n/\ln L_0\right)^{-2\ln L_0/N},$$

gdzie: L_n – maksymalna wiarygodność modelu zawierającego n zmiennych, L_0 – maksymalna wiarygodność modelu zawierającego jedynie wyraz wolny. Wszystkie wymienione miary przyjmują wartości z przedziału $[0,1]$[7].

Ocenę trafności prognoz można oprzeć na tzw. tablicy trafności, w której przewidywane $\hat{Y}$ szacuje się następująco: $\hat{Y}=\begin{cases}1\ dla\ \hat{p}>p^*\\0\ dla\ \hat{p}\le p^*\end{cases}$, gdzie p^* jest tzw. wartością odcinającą.

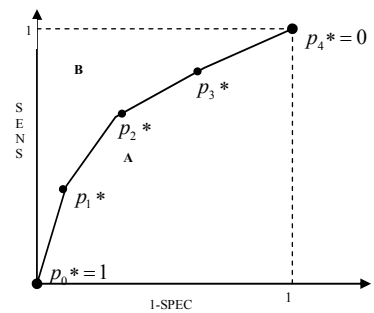
Zaobserwowane Y	Przewidywane $\hat{Y}$		Razem
	1	0	
1	n_{11}	n_{10}	$n_{1\bullet}$
0	n_{01}	n_{00}	$n_{0\bullet}$
Razem	$n_{\bullet 1}$	$n_{\bullet 0}$	N

Podstawowe miary zdolności predykcyjnej modelu to: zliczeniowy R^2 (*accurancy*) $R_z^2=(n_{00}+n_{11})/N$, czułość (*sensitivity*) $SENS = n_{11}/n_{1\bullet}$, swoistość (*specifity*) $SPEC = n_{00}/n_{0\bullet}$, dodatnia zdolność predykcyjna (*positive predictive value*) $PPV = n_{11}/n_{\bullet 1}$, ujemna zdolność predykcyjna[8] (*negative predictive value*) $NPV = n_{00}/n_{\bullet 0}$, wskaźnik dokładności dla krzywej koncentracji ROC (*receiver operating characteristic*) $WD=\frac{pole\ A}{pole\ A+pole\ B}$, gdzie A i B to figury zaznaczone na wykresie na rys. 2.

[6] Jest to unormowana miara Aldricha-Nelsona.

[7] Należy zwrócić uwagę, iż w modelach dychotomicznych niska wartość miary dopasowania, zwłaszcza przy dużych zbiorach danych, nie świadczy o złym dopasowaniu modelu [Gruszczyński 2001].

[8] Przy interpretacji dodatniej i ujemnej zdolności predykcyjnej należy pamiętać, że wartość tych miar zależy od proporcji jedynek w populacji.



Rys. 2. Estymacja krzywej koncentracji ROC
Źródło: opracowanie własne.

Tabela 1. Wyniki estymacji modelu dla semestru 1 dla kobiet

Model logitowy						
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*	iloraz szans e^b	95% przedział dla ilorazu szans
	–7,5161	1,4871	–5,0543	0,0000	0,0005	0,0000-0,0103
X_1	0,1314	0,0373	3,5233	0,0005	1,1404	1,0595-1,2275
X_2	0,7847	0,2499	3,1400	0,0020	2,1918	1,3383-3,5898
X_3	1,1904	0,3956	3,0091	0,0030	3,2885	1,5060-7,1808
X_4	3,0491	0,6863	4,4426	0,0000	21,0973	5,4429-81,7729
dla modelu: $\chi^2 = 37{,}585$; *p-value* = 0,0000001 kryterium informacyjne Akaikego AIC = 128,845 pseudo-R^2 Cragga-Uhlera = 32,8% pseudo-R^2 Vealla-Zimmermanna = 37,6%				zliczeniowy R^2 = 75,1% czułość modelu = 65,5% swoistość modelu = 77,1% wskaźnik dokładności = 0,670		
Model probitowy						
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*		
	– 4,2663	0,8002	– 5,3319	0,0000		
X_1	0,0737	0,0208	3,5474	0,0005		
X_2	0,4387	0,1359	3,2272	0,0015		
X_3	0,7013	0,2260	3,1027	0,0023		
X_4	1,7373	0,3882	4,4747	0,0000		
dla modelu: $\chi^2 = 38{,}118$; *p-value* = 0,0000001 kryterium informacyjne Akaikego AIC = 128,312 pseudo-R^2 Cragga-Uhlera = 33,2% pseudo-R^2 Vealla-Zimmermanna = 38,0%				zliczeniowy R^2 = 73,4% czułość modelu = 65,5% swoistość modelu = 75,0% wskaźnik dokładności = 0,657		

X_1 – wiek; X_2 – odległość miejsca zamieszkania podzielona według kwartyli na 4 kategorie (od 0 – najmniejsza do 3 − największa); X_3 – kierunek studiów: 0 *finanse i rachunkowość*, 1 *zarządzanie*, 2 *informatyka i ekonometria* (porządek według narastania ryzyka skreślenia);[9] X_4 – kontynuacja kierunku immatrykulacji: 0 − tak, 1 − nie.

Źródło: obliczenia własne przy pomocy programu STATISTICA 8.0.

[9] Nie poprawiła modelu zamiana zmiennej wielopoziomowej (odległość miejsca zamieszkania podzielona według kwartyli oraz kierunek studiów) na zmienne zero-jedynkowe odpowiadające poszczególnym klasom wartości przy ustaleniu jednej klasy jako poziomu odniesienia. Oznacza to, że iloraz szans jest taki sam dla sąsiadujących klas uporządkowanych według ryzyka.

Powyższe miary przyjmują wartości z przedziału $[0,1]$. Wartość miar zbudowanych na podstawie tablicy trafności zależy od przyjęcia punktu odcięcia p^*. Wybór p^* nie jest oczywisty, gdyż w miarę wzrostu wartości odcinającej p^* rośnie swoistość i jednocześnie maleje czułość modelu (wykres na rys. 2). W praktyce można zaproponować jedno z dwóch prostych kryteriów doboru p^*: przyjęcie za p^* proporcji jedynek w próbie lub wartości p^* dla punktu z wykresu ROC leżącego najbliżej punktu o współrzędnych (0,1). W badaniu za punkt odcięcia przyjęto proporcję jedynek[10]. Krzywa ROC zawiera więcej informacji niż tablica trafności, gdyż przedstawia siłę predykcji dla wszystkich możliwych punktów odcięcia p^* [Agresti 2002].

5. Porównanie modeli wynikowych: logitowego i probitowego

Dla kolejnych semestrów oszacowano prawdopodobieństwa skreślenia z listy studentów, pod warunkiem że student dotrwał do danego semestru (wyniki w tab. 1-4)[11].

Tabela 2. Wyniki estymacji modelu dla semestru 1 dla mężczyzn

Model logitowy						
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*	iloraz szans e^b	95% przedział dla ilorazu szans
	– 2,4716	0,6430	– 3,8440	0,0002	0,0844	0,0236-0,3016
X_1	1,2881	0,5900	2,1832	0,0310	3,6258	1,1275-11,6601
X_2	0,4824	0,2330	2,0701	0,0406	1,6200	1,0213-2,5697
dla modelu: $\chi^2 = 9{,}694$; *p-value* = 0,00786				zliczeniowy $R^2 = 60{,}5\%$		
kryterium informacyjne Akaikego AIC = 147,4604				czułość modelu = 64,9%		
pseudo-R^2 Cragga-Uhlera = 10,7%				swoistość modelu = 58,6%		
pseudo-R^2 Vealla-Zimmermanna = 13,2%				wskaźnik dokładności = 0,258		

Model probitowy				
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*
	– 1,5071	0,3654	– 4,1244	0,0001
X_1	0,7777	0,3329	2,3359	0,0211
X_2	0,2989	0,1385	2,1581	0,0329
dla modelu: $\chi^2 = 9{,}982$; *p-value* = 0,0068			zliczeniowy $R^2 = 60{,}5\%$	
kryterium informacyjne Akaikego AIC = 147,1716			czułość modelu = 64,9%	
pseudo-R^2 Cragga-Uhlera = 11,0%			swoistość modelu = 58,6%	
pseudo-R^2 Vealla-Zimmermanna = 13,6%			wskaźnik dokładności = 0,258	

X_1 – wiek dwie grupy: 0 dla osób mających ponad 26 lat, 1 dla osób mających do 26 lat (włącznie); X_2 – kierunek studiów: 0 *informatyka i ekonometria*, 1 *finanse i rachunkowość*, 2 *zarządzanie* (porządek według narastania ryzyka skreślenia).

Źródło: obliczenia własne przy pomocy programu STATISTICA 8.0.

[10] Drugie kryterium w tym przypadku dało zbliżony wynik.

[11] Ponieważ przetrwanie kobiet na pierwszym semestrze zależy od większej liczby czynników niż w przypadku mężczyzn, dla pierwszego semestru zostały oszacowane odrębne dla płci modele: logitowy oraz probitowy.

Tabela 3. Wyniki estymacji modelu dla semestru 2

Model logitowy						
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*	iloraz szans e^b	95% przedział dla ilorazu szans
	– 3,5581	0,5535	– 6,4285	0,0000	0,0285	0,0096-0,0848
X_1	1,2785	0,4585	2,7883	0,0057	3,5912	1,4550-8,8636
X_2	0,6311	0,2966	2,1276	0,0344	1,8796	1,0477-3,3721
dla modelu: χ^2 = 15,630; *p-value* = 0,0004 kryterium informacyjne Akaikego AIC = 148,738 pseudo-R^2 Cragga-Uhlera = 13,2% pseudo-R^2 Vealla-Zimmermanna = 15,6%				zliczeniowy R^2 = 70,1% czułość modelu = 60,0% swoistość modelu = 71,4% wskaźnik dokładności = 0,453		

Model probitowy				
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*
	– 1,9945	0,2684	– 7,4305	0,0000
X_1	0,6869	0,2361	2,9097	0,0040
X_2	0,3391	0,1486	2,2817	0,0234
dla modelu: χ^2 = 16,235; *p-value* = 0,0003 kryterium informacyjne Akaikego AIC = 148,133 pseudo-R^2 Cragga-Uhlera = 13,7% pseudo-R^2 Vealla-Zimmermanna = 16,1%			zliczeniowy R^2 = 70,1% czułość modelu = 60,0% swoistość modelu = 71,4% wskaźnik dokładności = 0,383	

X_1 – płeć: 0 dla kobiet, 1 dla mężczyzn; X_2 – kierunek studiów: 0 *finanse i rachunkowość*, 1 *informatyka i ekonometria*, 2 *zarządzanie* (porządek według narastania ryzyka skreślenia).

Źródło: obliczenia własne przy pomocy programu STATISTICA 8.0.

Tabela 4. Wyniki estymacji modelu dla semestru 3

Model logitowy						
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*	iloraz szans e^b	95% przedział dla ilorazu szans
	– 2,9220	0,4451	– 6,5642	0,0000	0,0538	0,0224-0,1295
X_1	0,8406	0,2688	3,1270	0,0020	2,3178	1,3642-3,9379
dla modelu: 11,561; *p-value* = 0,00067 kryterium informacyjne Akaikego AIC = 152,465 pseudo-R^2 Cragga-Uhlera = 10,1% pseudo-R^2 Vealla-Zimmermanna = 12,2%				zliczeniowy R^2 = 64,6% czułość modelu = 66,7% swoistość modelu = 64,2% wskaźnik dokładności = 0,367		

Model probitowy				
Stała	ocena parametru b_i	błąd standardowy	statystyka t	*p-value*
	– 1,6466	0,2116	– 7,7808	0,0000
X_1	0,4437	0,1360	3,2627	0,0013
dla modelu: 11,632; *p-value* = 0,0006 kryterium informacyjne Akaika AIC = 152,394 pseudo-R^2 Cragga-Uhlera = 10,2% pseudo-R^2 Vealla-Zimmermanna = 12,2%			zliczeniowy R^2 = 64,6% czułość modelu = 66,7% swoistość modelu = 64,2% wskaźnik dokładności = 0,367	

X_1 – kierunek studiów: 0 *finanse i rachunkowość*, 1 *informatyka i ekonometria*, 2 *zarządzanie* (porządek według narastania ryzyka skreślenia).

Źródło: obliczenia własne przy pomocy programu STATISTICA 8.0.

Tabela 5. Tablica odejść studentów ze studiów (studia niestacjonarne 3,5-letnie)

Semestr	Liczba rozpoczynających	Liczba cenzurowanych	Liczba zagrożonych	Liczba skreśleń	Estymator aktuarialny	Estymator proporcji [12]
1	297	0	297	66	0,222222	0,222222
2	231	0	231	25	0,108225	0,108225
3	206	4	204	27	0,132353	0,131068
4	175	4	173	7	0,040462	0,040000
5	164	1	164	4	0,024465	0,024390
6	159	12	153	1	0,006536	0,006289
7	146	146	73	0	0	0

Źródło: opracowanie własne.

W odniesieniu do następnych semestrów nie udało się wyodrębnić w modelu statystycznie istotnych czynników objaśniających, więc za estymator prawdopodobieństwa zdarzenia przyjęto proporcję z próby (tab. 5). Ze względu na występowanie danych cenzurowanych prawdopodobieństwo zdarzenia może być oszacowane za pomocą estymatora aktuarialnego, który ma postać: $\hat{p}_k = d_k / (l_k - 0,5c_k)$, gdzie l_k – liczba osób rozpoczynających *k*-ty semestr, d_k – liczba studentów skreślonych z listy w czasie semestru, c_k – liczba studentów, którzy nie ukończyli semestru, ale ciągle studiują[13].

6. Podsumowanie

Dystrybuanty rozkładu logistycznego i normalnego są zbliżone, więc modele logitowy i probitowy dają podobne wyniki. Ocena parametru b_i modelu logitowego jest porównywalna z oceną parametru b_i modelu probitowego przemnożonego przez $\pi/\sqrt{3}$ (odchylenie standardowe rozkładu logistycznego). Zaletą modelu logitowego jest interpretacja parametrów przy wykorzystaniu ilorazu szans. Porównania modeli można dokonać na dwa sposoby: za pomocą miar dopasowania typu R^2 lub za pomocą miar trafności prognoz. Przy wyborze zmiennych do modelu można korzystać z kryteriów opartych na wartości funkcji wiarygodności: testu chi-kwadrat do oceny dopasowania modelu, jak też kryterium informacyjnego Akaikego − AIC [Agresti 2002; Harrell 2006].

[12] W modelu logitowym średnie prawdopodobieństwo oszacowane na podstawie próby jest równe proporcji wystąpienia zdarzenia w próbie [Gruszczyński 2001, s. 63].

[13] W analizie przeżycia, oprócz estymatora aktuarialnego, często stosuje się estymator Kaplana-Meiera. Estymator ten w przypadku skończonej liczby momentów, w których może dojść do zdarzenia, oraz dużej liczby narażonych jednostek (w każdym momencie dochodzi co najmniej do jednego zdarzenia) jest równoważny estymatorowi proporcji $\hat{p}_k = d_k / l_k$ o rozkładzie $B(p, \sqrt{p(1-p)/N})$.

Narzędzia regresji zarówno logistycznej, jak i probitowej pozwoliły na wykrycie prawidłowości w procesie odejść ze studiów w badanej kohorcie. Na semestrze pierwszym ryzyko jest wyższe dla mężczyzn niż dla kobiet. Najbardziej zagrożeni są młodzi mężczyźni studiujący na kierunku *zarządzanie*. Natomiast wśród kobiet ryzyko rośnie wraz z wiekiem oraz z odległością miejsca zamieszkania od uczelni. Największe ryzyko dla kobiet występuje na kierunku *informatyka i ekonometria*. Dla kobiet, które rozpoczęły wcześniej studia na innym kierunku, ryzyko jest większe ponad 21-krotnie. Na semestrze drugim istotnymi predyktorami przerwania studiów są płeć i kierunek studiów, z tym że największe ryzyko dotyczy kierunku *zarządzanie*. Na semestrze trzecim istotnym czynnikiem jest już tylko kierunek studiów. Studenci kierunku *Zarządzanie* są przeszło dwukrotnie bardziej narażeni niż studenci *informatyki i ekonometrii* i przeszło pięciokrotnie bardziej niż studenci kierunku *finanse i rachunkowość*. Dla semestrów 4-7 nie zidentyfikowano istotnych zmiennych objaśniających. W celu prospektywnej weryfikacji modelu badaniem należałoby objąć następne kohorty studentów.

Literatura

Agresti A.(2002), *Categorical data analysis*, Wiley-Interscience, New Jersey.

Gruszczyński M. (2001), *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, SGH, Warszawa.

Harrell F. (2001), *Regression modeling strategies whit applications to linear models, logistic regression, and survival analysis*, Springer-Verlag, New York.

Williams B.A., Mandrekar J.M., Cha S.S., Furth A.F. (2006), *Finding optimal cutpoints for continuous covariates with binary and time-to-event outcomes*, „Technical Report", Mayo Foundation, Rochester.

MODELS OF RISK RELEGATION BASED ON THE EXAMPLE OF EXTRAMURAL STUDENTS

Summary

On the basis of observations made on the 2004/2005 cohort of extramural students of the Faculty of Management, University of Gdansk, there was made an attempt to identify risk factors of relegation. In order to estimate probabilities of relegation during semesters, we used logit and probit models. The significance levels were examined with the use of chi-square test and Akaike information criterion. As a goodness of fit criteria were used pseudo-R^2 measures and accuracy coefficients (based on ROC curve). The probabilities of rejection for these semesters were also estimated with the use of actuarial estimator of proportion due to presence of censored data.