PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1105 2006
Zastosowanie statystyki w ekonomii

Joanna Zdobylak
Akademia Ekonomiczna we Wrocławiu

IDENTYFIKACJA OBSERWACJI ODSTAJĄCYCH W BADANIACH ANKIETOWYCH Z WYKORZYSTANIEM MODELI REGRESJI BINARNYCH

1. Wstęp

Współczesne czasy często określane są mianem ery informacji. To właśnie informacja jest czynnikiem, który decyduje o rozwoju i konkurencyjności małego przedsiębiorstwa lub też całego kraju. Dlatego posiadanie odpowiednich informacji jest niezwykle ważne. Istnieje wiele źródeł uzyskiwania informacji i wiele sposobów ich pozyskiwania. Jednym z najczęściej stosowanych są badania społeczno-ekonomiczne. W badaniach społeczno-ekonomicznych wykorzystuje się różne metody i techniki badawcze. Do metod badawczych zalicza się np. eksperymenty laboratoryjne, obserwacje systematyczne. Wśród technik badawczych wyróżnia się m.in. ankiety, wywiady, spisy czy też innego rodzaju techniki socjometryczne.

Podczas etapu zbierania danych z wykorzystaniem wymienionych metod i technik bardzo często występuje problem obserwacji odstających (ang. *outliers*). Przyczyny występowania tego problemu są różne, np. brak chęci udzielenia przez respondenta poprawnej odpowiedzi na stawiane mu pytania lub też występowanie w badanej grupie jednostki nietypowej (np. wybitnie uzdolnionej). Można mieć również do czynienia z niejednorodną próbą statystyczną, z której próbka została pobrana (pochodzą z innej grupy społecznej), bądź też z błędem popełnionym przez badacza (np. podczas robienia pomiaru, kodowania wyników). Dane odstające postrzegane są jako dane, które zniekształcają informacje uzyskane podczas badania, np. zmieniają i wypaczają charakter zależności między badanymi cechami. Istnieje wiele skal pomiarowych, jednak w badaniach społeczno-ekonomicznych najczęściej są stosowane cztery spośród nich: nominalna, porządkowa, przedziałowa i ilorazowa.

W celu przeciwdziałania niedogodnościom związanym z występowaniem obserwacji odstających wykorzystuje się metody identyfikacji danych odstających.

Celem artykułu jest przedstawienie zastosowania metod wykrywania obserwacji odstających do analizy danych pochodzących z badań ankietowych. Zostanie przeprowadzona analiza otrzymanych wyników i podane będą praktyczne wskazówki dla osób chcących stosować przedstawione metody.

Dane odstające są zazwyczaj definiowane jako obserwacje niepochodzące z wcześniej założonego modelu albo jako wartości, które są znacznie oddalone od pozostałych obserwacji. Trudno jest jednak zdefiniować, czym są obserwacje odstające, ponieważ pojęcie obserwacji znacznie oddalonych od pozostałych jest względne. Autorka zatem nie poda uniwersalnej definicji, poda natomiast definicję, która będzie dostosowana do rozważanego problemu oraz przyjętego modelu.

W badaniach ankietowych zwykle jest tak, że zjawiska, o które pytamy, mają charakter jakościowy i zmienne je opisujące są zmiennymi dyskretnymi. Z reguły ma się do czynienia z nimi wtedy, gdy dane dotyczą pewnych jednostek, np. osób, które mają dokonać wyboru spośród różnych możliwości. Przykładowo odpowiedzi dotyczące posiadania samochodu lub nie, wykształcenia (podstawowe, zawodowe, średnie, wyższe) itp. Wybór odpowiedzi jest w pewnym stopniu uzależniony od różnych czynników, będących w tym przypadku zmiennymi objaśniającymi, np. podczas rozważania posiadania samochodu bądź nie takim czynnikiem jest niewątpliwie dochód. W artykule rozważone zostaną tylko modele z binarnymi zmiennymi objaśniającymi (tzn. przyjmującymi jedną z dwu wartości: 0 lub 1). Wyjaśniają one prawdopodobieństwo zajścia jednej z możliwości, w zależności od czynników, które warunkują to zajście (por. [*Estymacja modeli...* 1990; Domański, Pruska 2002]).

W pracy proponowanym narzędziem wykrywania obserwacji odstających jest diagnostyka reszt. Reszty odzwierciedlają niezgodność pomiędzy wartościami obserwowanymi i przewidywanymi przez model. Jeżeli model jest dobrze dobrany, wartości reszt powinny być „małe". Obserwacje nietypowe często cechują się dużymi wartościami reszt. Jeżeli tak nie jest, taki punkt odstający jest nie do wykrycia.

2. Zjawiska opisywane za pomocą zmiennych binarnych i metody wykrywania obserwacji odstających w takich sytuacjach

W pracy rozważone zostaną metody wykrywania obserwacji odstających w sytuacjach, gdy objaśniana zmienna kategorialna **Y** jest binarna (możliwe odpowiedzi: tak lub nie). Dane tego typu spotyka się często w ankietach. Załóżmy, że zmienna objaśniana **Y** zależy od k zmiennych objaśniających.

Metody wykrywania obserwacji odstających

W praktyce stosuje się wiele kryteriów umożliwiających wykrywanie obserwacji odstających. W pracy zostaną omówione następujące metody (por. [Barnett, Lewis 1994; Johnson, Albert 1999]):

- reszty Pearsona (reszty standaryzowane) $r_{i,P}$,

- reszty odchyleń $r_{i,D}$,
- dopasowane reszty odchyleń $r_{i,AD}$.

Reszty Pearsona

Reszty Pearsona dla modelu regresji binarnej są definiowane następująco:

$$r_{i,P} = \frac{y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}, \tag{1}$$

gdzie n_i jest liczebnością danej klasy.

Gdy wyznaczamy obserwacje odstające w badaniach ankietowych $n_i = 1$, ponieważ n_i w tym przypadku to respondent, który wypełnił ankietę (tzn. każda z klas jest jednoelementowa).

Reszty Pearsona najlepiej stosować dla $n_i \geq 5$ albo, ściślej mówiąc, gdy $n_i p_i \geq 3$ i $n_i(1 - p_i) \geq 3$, ponieważ wówczas są one dosyć dobrze przybliżane przez standaryzowany rozkład normalny.

Reszty odchyleń $r_{i,D}$

Reszty odchyleń odnoszą się do funkcji ilorazu wiarygodności D, która ma następującą postać:

$$D = 2\sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right\}.$$

Statystyka ta ma rozkład chi-kwadrat z $n - k - 1$ stopniami swobody, gdzie k jest liczbą zmiennych objaśniających.

Reszty odchyleń $r_{i,D}$ są określane dla i-tej obserwacji jako:

$$r_{i,D} = \text{sign}(y_i - \hat{y}_i) \left\{ 2 \left[y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right] \right\}^{1/2}, \tag{2}$$

gdzie przewidywane wartości $\hat{y}_i = n_i \hat{p}_i$, sign() ma wartość 1 lub (–1), jeśli argument ma wartość odpowiednio dodatnią albo ujemną. Podobnie jak reszty Pearsona, reszty odchyleń mają rozkład normalny, gdy $n_i p_i \geq 3$ i $n_i(1 - p_i) \geq 3$.

Reszty Pearsona oraz reszty odchyleń należy stosować ostrożnie, gdy przewidywane wartości są mniejsze od 3, gdyż wówczas ich rozkład znacznie różni się od rozkładu normalnego. Mimo że reszty Pearsona oraz reszty odchyleń mają asymptotycznie ten sam rozkład, to mogą dać różne wyniki.

Dopasowane reszty odchyleń $r_{i,AD}$

Dopasowane reszty odchyleń są uzyskiwane z reszt odchyleń dzięki zastosowaniu poprawki pierwszego rzędu do obciążenia średniej. Zatem dla modelu regresji binarnej reszty mają następującą postać:

$$r_{i,AD} = r_{i,D} + \frac{1-2\hat{p}_i}{6\sqrt{n_i \hat{p}_i (1-\hat{p}_i)}}. \tag{3}$$

Z tych trzech proponowanych reszt rozkład dopasowanych reszt odchyleń jest najbliższy rozkładowi normalnemu i daje zaskakująco dokładne przybliżenie nawet dla małych n_i $(n_i p_i \leq 3)$, przy założeniu, że szacowane prawdopodobieństwo nie jest zbyt bliskie 0 albo 1.

Obserwacje uznaje się za odstające, gdy wartości reszt są co do wartości bezwzględnej większe bądź bardzo bliskie 3.

Aby można było wykorzystać przedstawione metody do wykrywania obserwacji odstających, należy dla każdego *i*-tego respondenta oszacować prawdopodobieństwa sukcesu p_i zajścia badanego zdarzenia *E*, przy założeniu, że stan zmiennych objaśniających dla *i*-tego respondenta jest $\mathbf{X} = [x_{i1}, x_{i2}, \ldots, x_{ik}]$.

Najprostszym modelem, który można zastosować, jest tzw. **liniowy model** prawdopodobieństwa. Zakłada się w nim, że prawdopodobieństwo sukcesu zajścia zdarzenia *E* dla *i*-tego respondenta może być wyrażone jako:

$$p_i = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_k x_{ik} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}. \tag{4}$$

W tym modelu zakłada się również, że prawdopodobieństwo sukcesu p_i zmienia się liniowo wraz z wartościami danych $\mathbf{X}$. Do estymacji współczynników $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_k$, nazywanych współczynnikami regresji, stosuje się metodę najmniejszych kwadratów.

Dopasowanie modelu regresji liniowej do danych może doprowadzić do sytuacji, że wartości teoretyczne prawdopodobieństwa p_i, otrzymane dla modelu (4), znajdują się poza przedziałem $\langle 0,1 \rangle$. Powoduje to zatem konieczność poszukiwania modeli innych typów.

Jednym z rozwiązań tego problemu jest zastosowanie do estymacji prawdopodobieństwa sukcesu p_i funkcji, która daje wartości tylko z przedziału $\langle 0, 1 \rangle$. W odniesieniu do danych statystycznych klasą takich funkcji jest dystrybuanta. Wprowadzając dystrybuantę w relację pomiędzy prawdopodobieństwem sukcesu p_i a zmiennymi objaśniającymi $\mathbf{X}$, modyfikuje się model (4) przez założenie, że:

$$p_i = F(\beta_0 + \beta_1 \cdot x_{i1} + ... + \beta_k x_{ik}) = F(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}), \tag{5}$$

gdzie F jest dystrybuantą, wybór funkcji F jest zaś dowolny (por. [*Estymacja modeli...* 1990]).

2.1. Modele regresji binarnych

W klasycznych metodach dopasowania modelu do danych proponuje się kilka funkcji, a ich wybór jest zależny od najlepszego dopasowania do danych.

W zależności od przyjętych postaci rozkładu F otrzymuje się różne modele. Do danych binarnych stosuje się następujące modele (por. [Barnett, Lewis 1994; *Estymacja modeli...* 1990; Johnson, Albert 1999]):

1. **Model probitowy**

$$p_i = \Phi(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}), \tag{6}$$

gdzie Φ oznacza dystrybuantę standaryzowanego rozkładu normalnego.

2. **Model logitowy**, który jest wyprowadzany z dystrybuanty rozkładu logistycznego, wyrażonej w następujący sposób:

$$F(x) = \frac{e^x}{1+e^x}, \qquad -\infty < x < +\infty.$$

Wyrażenie prawdopodobieństwa sukcesu w formie liniowej ma następującą postać:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}. \tag{7}$$

Wielkość $\ln(p_i/(1-p_i))$ nazywana jest przekształceniem logistycznym prawdopodobieństwa sukcesu p_i albo krócej – logit. Prawdopodobieństwo sukcesu p_i ma zatem postać:

$$p_i = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}. \tag{8}$$

Rozkład logistyczny jest symetryczny względem punktu 0, ma kształt zbieżny z rozkładem normalnym, ma cięższe ogony oraz jest bardziej stromy (por. [Jennings 1986]).

3. **Model log-log liniowy.**

Inną powszechnie używaną funkcją jest bazujący na wartościach ekstremalnych rozkład

$$F(x) = 1 - \exp(-\exp(x)) \qquad -\infty < x < +\infty .$$

Wykorzystując postać dystrybuanty, otrzymuje się zatem model prawdopodobieństwa sukcesu p_i w następującej postaci (formie):

$$\ln[-\ln(1-p_i)] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}. \tag{9}$$

Funkcja łącząca w tym modelu to log-log dopełnienie.

2.2. Estymacja parametrów β modeli regresji binarnych

Kolejnym etapem jest estymacja parametrów modeli regresji binarnych.

Do estymacji parametrów modeli regresji, a co się z tym wiąże – prawdopodobieństwa sukcesu p_i – wykorzystuje się dwa źródła: próbę statystyczną oraz informacje, które pochodzą spoza próby, tzw. informacje *a priori* o parametrach. Estymacja klasyczna wykorzystuje tylko informacje zawarte w próbie i przyjmuje, że parametr jest pewną wartością stałą. Natomiast stosując estymację nieklasyczną, zwaną bayesowską, oprócz wiadomości zawartych w próbie statystycznej, korzysta się również z informacji *a priori* (por. [DeGroot 1981]).

2.2.1. Metoda estymacji parametrów β (metoda największej wiarygodności)

W podejściu klasycznym do estymacji parametrów przyjętego modelu korzysta się z informacji pochodzących tylko z próby statystycznej. Ponadto zakłada się, że szacowane parametry są pewną nieznaną wartością stałą. Do estymacji parametrów $\boldsymbol{\beta}$ stosowana jest metoda największej wiarygodności. Chcąc wykorzystać metodę największej wiarygodności do estymacji parametrów $\boldsymbol{\beta}$, należy podać postać funkcji wiarygodności dla parametrów $\boldsymbol{\beta}$.

Przypomnijmy, że w omawianym przypadku badana zmienna **Y** ma rozkład zero-jedynkowy o funkcji prawdopodobieństwa o postaci:

$$f(y_i|p_i) \propto p_i^{y_i}(1-p_i)^{n_i-y_i}, \; y_i \in \{0, 1\}, \; 0 < p_i < 1, \; N = \sum_{i=1}^{n} n_i .$$

Funkcja wiarygodności ma zatem postać:

$$L(p) \propto \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{n_i-y_i} \text{, gdzie } N = \sum_{i=1}^{n} n_i, \; n_i = 1 .$$

Dla N-niezależnych obserwacji z prawdopodobieństwem sukcesu p_i, wyznaczonym jednym z powyższych metod, funkcję wiarygodności parametru $\boldsymbol{\beta}$ można przedstawić następująco:

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{N} F(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})^{y_i} (1 - F(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}))^{n_i - y_i} .$$

Do praktycznego zastosowania tej metody niezbędne jest korzystanie z właściwego algorytmu iteracyjnego, ponieważ nie można analitycznie wyznaczyć ekstremum funkcji $L(\boldsymbol{\beta})$. Obecnie jednak algorytmy iteracyjne są zamieszczane w pakietach statystycznych, np. program Statistica (por. [Angresti 1990; Johnson, Albert 1999]).

2.2.2. Bayesowska metoda estymacji parametrów $\boldsymbol{\beta}$

W przypadku stosowania nieklasycznego modelu estymacji parametrów korzysta się, jak powiedziano wcześniej, z informacji pochodzących spoza próby o badanym zjawisku, tzw. informacji *a priori*. Zakłada się, że szacowane parametry są zmienną losową. Zatem, aby stosować ten rodzaj estymacji, należy mieć wstępne informacje o estymowanym parametrze, np. na podstawie wstępnej próby. Będziemy korzystać z podejścia, w którym mamy tzw. rozkład *a priori* tego parametru stworzony na podstawie informacji pochodzących spoza próby.

Do wyznaczania wartości parametrów $\boldsymbol{\beta}$ dla modelu regresji binarnej w estymacji bayesowskiej proponowany jest algorytm Metropolisa-Hastingsa. Jest to iteracyjna technika, polegającą na tworzeniu próbek pseudolosowych. Próbki pseudolosowe są generowane ze wspólnego rozkładu *a posteriori* dla parametrów regresji $\boldsymbol{\beta}$. Jako wartości startowej algorytm ten używa wartości estymatorów największej wiarygodności $\hat{\boldsymbol{\beta}}$ i macierzy kowariancji $\hat{\mathbf{C}}$. Dokładnie algorytm jest podany w [Jennings 1986].

3. Przykład zastosowania omówionych metod

Prezentowany przykład ilustruje zastosowanie metod wykrywania obserwacji, odstających wykorzystujących klasyczną estymację parametrów $\boldsymbol{\beta}$.

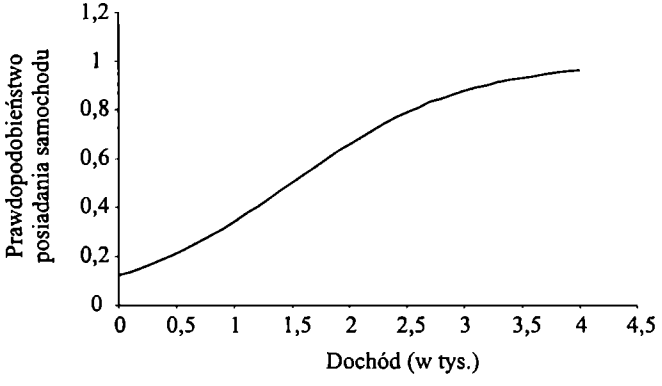
W wyniku badań ankietowych przeprowadzonych na losowo wybranych osobach pracujących w urzędach w powiecie lubańskim uzyskano dane dotyczące sytuacji materialnej rodzin oraz posiadania przez badane rodziny samochodu. Uzyskano dane dotyczące 45 gospodarstw domowych, pochodzących z jednej warstwy społecznej. Dane dotyczące dochodów potraktowano jako realizację zmiennej losowej objaśniającej $\mathbf{X}$, natomiast dane dotyczące posiadania samochodu – jako realizację zmiennej losowej objaśnianej $\mathbf{Y}$. Możliwe realizacje zmiennej losowej $\mathbf{Y}$ to 0, gdy badana rodzina nie ma samochodu, oraz 1, gdy rodzina ma samochód. Uzyskane odpowiedzi zawarte zostały w tab. 1.

Tabela 1. Dane dotyczące sytuacji materialnej badanych rodzin oraz posiadania przez nie samochodu

Lp.	Dochód (w tys. zł)	Samochód	Lp.	Dochód (w tys. zł)	Samochód	Lp.	Dochód (w tys. zł)	Samochód
1	3,20	1	16	2,00	1	31	1,50	0
2	1,00	0	17	1,38	0	32	3,00	1
3	4,00	1	18	1,80	0	33	1,03	0
4	1,99	1	19	2,50	1	34	1,70	1
5	2,00	0	20	1,24	0	35	3,40	1
6	1,82	1	21	1,60	1	36	2,70	1
7	2,00	1	22	1,35	0	37	1,90	1
8	2,00	1	23	1,16	0	38	0,98	0
9	4,80	0	24	0,95	0	39	1,20	0
10	2,00	1	25	1,90	1	40	0,82	1
11	3,30	1	26	1,16	0	41	2,10	1
12	2,00	1	27	1,05	0	42	3,30	1
13	1,50	1	28	1,50	0	43	0,75	1
14	2,40	1	29	1,20	0	44	1,30	1
15	3,00	1	30	1,50	0	45	2,70	0

Źródło: niepublikowane dane pochodzące z badania przeprowadzonego przez PCE w Lubaniu.

Na podstawie danych empirycznych oszacowano parametry β modeli regresji binarnych, wykorzystując metodę największej wiarygodności. Okazało się, że do danych tych dobrze dopasowany jest model logitowy ($p - value = 0{,}0097$). Wykres prawdopodobieństwa posiadania samochodu przez badane rodziny uzyskany za pomocą modelu logitowego przedstawiono na rys. 1.



Rys. 1. Wykres prawdopodobieństwa posiadania samochodu
Źródło: wykonano w programie Statistica.

Tabela 2. Wartości reszt dla dopasowanego modelu logitowego

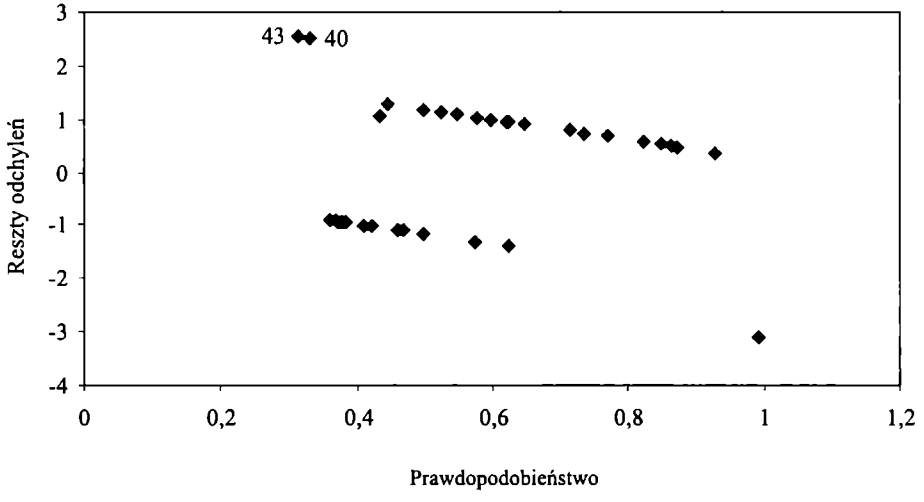
Lp.	p_i	$r_{i,P}$	$r_{i,D}$	$r_{i,AD}$
1	0,849	0,42	0,57	0,53
2	0,372	–0,77	–0,96	–0,94
3	0,927	0,28	0,39	0,35
4	0,620	0,78	0,97	0,95
5	0,623	–1,28	–1,39	–1,41
6	0,578	0,85	1,05	1,03
7	0,623	0,78	0,97	0,95
8	0,623	0,78	0,97	0,95
9	0,992	**–11,13**	**–3,11**	**–3,12**
10	0,623	0,78	0,97	0,95
11	0,862	0,4	0,55	0,51
12	0,623	0,78	0,97	0,95
13	0,497	1,00	1,18	1,18
14	0,713	0,63	0,83	0,79
15	0,821	0,46	0,62	0,57
16	0,623	0,78	0,97	0,95
17	0,467	–0,93	–1,12	–1,11
18	0,573	–1,16	–1,30	–1,31
19	0,734	0,60	0,78	0,74
20	0,432	–0,87	–1,06	1,05
21	0,523	0,95	1,14	1,13
22	0,459	–0,92	–1,1	–1,09
23	0,411	–0,83	–1,03	–1,01
24	0,360	–0,75	–0,94	–0,91
25	0,598	0,82	1,01	0,99
26	0,411	–0,83	–1,03	–1,01
27	0,384	–0,79	–0,98	–0,96
28	0,497	–0,99	–1,17	–1,16
29	0,421	–0,85	–1,04	–1,02
30	0,497	–0,99	–1,17	–1,16
31	0,497	–0,99	–1,17	–1,16
32	0,821	0,46	0,63	0,58
33	0,379	–0,78	–0,97	–0,95
34	0,548	0,91	1,1	1,09
35	0,873	0,38	0,52	0,47
36	0,771	0,55	0,72	0,68
37	0,598	0,82	1,01	0,99
38	0,368	–0,76	–0,95	–0,92
39	0,421	–0,85	–1,04	–1,02
40	0,330	**2,42**	**2,49**	**2,51**
41	0,646	0,74	0,93	0,90
42	0,862	0,4	0,54	0,49
43	0,314	**2,48**	**2,52**	**2,55**
44	0,446	1,11	1,27	1,27
45	0,372	–0,77	–0,96	–0,93

Źródło: opracowanie własne.

Następnie na podstawie wzoru (8) wyznaczono wartości prawdopodobieństwa sukcesu p_i dla każdej z badanych osób. W kolejnym kroku dla wyznaczonych wartości prawdopodobieństw wyznaczono reszty Pearsona, reszty odchyleń $r_{i,D}$, dopasowane reszty odchyleń $r_{i,AD}$. Wyniki zamieszczono w tab. 2.

Na podstawie uzyskanych wyników można stwierdzić, że w próbie jest na pewno jedna osoba, której wartość reszty jest większa niż 3 (osoba 9). Osobę tę można uznać za niepasującą do modelu, czyli odstającą. Należy się jednak również zastanowić, czy osoby o numerach 40 i 43 nie są również podejrzane o niepasowanie do modelu. Najlepiej jest sporządzić wykres reszt (zob. rys. 2) i zobaczyć, jak rozproszyły się ich wartości, i dopiero wówczas należy wyciągnąć wnioski. Z rysunku wynika, że osoby 40 i 43 również można potraktować jako niepasujące do modelu.

Z prezentowanego przykładu widać, że należy sprawdzać jakość danych statystycznych i w późniejszej analizie zastanowić się, czy to, że osoby nie pasują do modelu, nie jest spowodowane jakimiś czynnikami. Na przykład w przedstawionym przykładzie osoba 9, która zarabia 4800 zł i nie ma samochodu, może nie mieć również prawa jazdy i samochód jest jej niepotrzebny, albo osoba 40 ma samochód, bo wygrała go w jakimś konkursie lub dostała w prezencie. Wyciągając zatem wnioski, należy również podać możliwe przyczyny występowania obserwacji odstających.



Rys. 2. Zestawienie reszt odchyleń z dopasowanymi wartościami prawdopodobieństw regresji logitowej

Źródło: wykonano w programie Statistica.

Może się jednak zdarzyć, że stosowanie różnych reszt może dać odmienne wyniki, dlatego wyciągając wnioski co do występowania obserwacji odstających,

najlepiej zastosować więcej niż jedną metodę i jako obserwacje odstające traktować te, które zostaną wykryte przez więcej niż jedną metodę.

4. Podsumowanie

Przedstawione metody pozwalają na rozwiązanie problemu znajdowania obserwacji niepasujących do rozważanego modelu. Podstawą oceny obserwacji jako odstających jest analiza merytoryczna. Należy podkreślić, iż gruntowna znajomość badanego zjawiska powinna być uzupełniona analizą statystyczną.

Metody: reszty Pearsona oraz reszty odchyleń powinny być stosowane do badań przeprowadzanych na dużą skalę. Zastosowanie tych dwóch metod w prezentowanym przykładzie miało tylko pokazać, jak stosować te metody do analizy danych.

Powszechna dostępność do oprogramowania umożliwiającego estymację jest zaletą przedstawionych metod i sprawia, że ich stosowanie jest wygodnym narzędziem do wykrywania obserwacji odstających.

Należy również pamiętać, że liczba obserwacji odstających stanowi niewielki odsetek danych. Gdy liczba obserwacji odstających jest znaczna, zazwyczaj ma się do czynienia z danymi niejednorodnymi. Problem w takiej sytuacji może zostać rozwiązany np. za pomocą odpowiednich metod klasyfikacji.

Problemy związane z występowaniem obserwacji odstających, które komplikują proces wyciągania wniosków podczas przeprowadzanej analizy, są do końca nie rozwiązane, zwłaszcza w badaniach społeczno-ekonomicznych. Zatem zawsze po dopasowaniu i oszacowaniu parametrów modelu do danych powinno się przeprowadzić analizę reszt i statystyk z nimi związanych, gdyż gwarantuje to wykrycie obserwacji odstających i wszystkich odstępstw oraz ich odpowiednią późniejszą analizę.

Literatura

Angresti A., *Categorial Data Analysis*, J.Wiley & Sons, New York 1990.
Barnett V., Lewis T., *Outliers in statistical data*, 3 wyd. J. Wiley & Sons, New York 1994.
Estymacja modeli ekonometrycznych, red. S. Bartosiewicz, PWE, Warszawa 1990.
DeGroot M., *Optymalne decyzje statystyczne*, PWN, Warszawa 1981.
Domański C., *Testy statystyczne*, PWN, Warszawa 1990.
Domański C., Pruska K., *Nieklasyczne metody statystyczne*, PWE, Warszawa 2000.
Everitt B. S., Dunn G., *Applied Multivariate Data Analysis*, Hodder & Stoughton, London 1991.
Johnson V.E., Albert J.H., *Ordinal Data Modeling*, Springer, New York 1999.
Jennings E.D., *Outliers and Residual Distributions in Logistic Regression*, JASA, 81, 987-990, 1986.

THE IDENTIFICATION OF OUTLIERS IN SURVEY RESEARCH USING BINARY REGRESSION MODELS

Summary

One of the most popular ways to get information about socio-economic occurrences is statistical research that uses questionnaires as a research tool. Within the analysis of this type of data the problem of outliers often appears. This article presents some of the methods of outliers identification. Those methods use binary regression. The author illustrates presented methods using an empirical example that shows advantages and disadvantages of those methods.