PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Grażyna Dehnel
Akademia Ekonomiczna w Poznaniu

ESTYMACJA WINSORA W BADANIACH PODMIOTÓW GOSPODARCZYCH

1. Wstęp

Zmienne w statystyce gospodarczej charakteryzują się, m.in. ze względu na obecność obserwacji odstających oraz znaczny odsetek jednostek z zerowymi wartościami cech, silną prawostronną asymetrią. Stąd też stosowanie klasycznych estymatorów jest niewłaściwe. Często bowiem nie zachowują swoich własności, takich jak nieobciążoność czy duża efektywność.

Wciąż poszukuje się nowych rozwiązań, które prowadziłyby do otrzymania wiarygodnych szacunków w przekroju małych domen czy dla małych obszarów.

Jedno z proponowanych w literaturze podejść do estymacji stosowanej w statystyce gospodarczej [Kokic, Bell 1994; Chambers 1996] polega na dokonaniu pewnej modyfikacji, która ma na celu „uodpornienie" estymatora na duże reszty. Jednostki wylosowane do próby, u których wartości cechy wykraczają poza pewne ustalone punkty graniczne, zostają zmienione. Ta zmiana może odbywać się poprzez zmodyfikowanie wartości wag związanych z obserwacjami odstającymi lub poprzez zmodyfikowanie wartości zmiennych tych jednostek.

Przykładem tego drugiego podejścia jest estymacja typu Winsora [Searls 1966]. Polega ona na dokonaniu podziału jednostek wylosowanych do próby na dwie grupy. Jedną z nich stanowią dane wykorzystane do budowy modelu, drugą – obserwacje odstające. Rozdziału dokonuje się na podstawie wyznaczonych wcześniej dwóch punktów granicznych. Następnie wartości badanej zmiennej jednostek znajdujących się poza punktami granicznymi przekształcane są tak, by nie stanowiły obserwacji odstających. Należy jednak podkreślić, że zmodyfikowane wartości zmiennej badanej są sztuczne.

Dalsze obliczenia, w których można się posłużyć dowolnym rodzajem estymacji stosowanej w badaniach częściowych, jak estymacja typu Horvitza–Thompsona czy GREG, prowadzone są na podstawie zmodyfikowanej próby.

Zaletą estymacji Winsora jest to, że dąży ona do minimalizacji błędu średniokwadratowego (MSE) nawet kosztem dużego obciążenia [Hedlin 2004]. Największą trudność stanowi wyznaczenie odpowiednich punktów granicznych pozwalających na dokonanie podziału jednostek. Właściwy ich dobór w znacznym stopniu poprawia jakość szacunku. Jedna z metod wykorzystywanych do określenia punktów granicznych wymaga oszacowania wartości obciążenia oraz wyznaczenia parametrów regresji [Kokic, Bell 1994]. W tym celu wykorzystuje się np. techniki regresji odpornej.

W niniejszym artykule podjęto próbę empirycznej weryfikacji możliwości wykorzystania estymacji Winsora do szacowania informacji o działalności gospodarczej małych przedsiębiorstw w przekroju małych domen (tj. województw i sekcji PKD). Celem badania była ocena wpływu wyboru technik regresji odpornej stosowanej przy klasyfikacji podmiotów gospodarczych na wartości punktów granicznych w estymacji Winsora.

2. Estymacja Winsora – podstawy teoretyczne

Estymator Winsora typu II jest dany wzorem:

$$\hat{Y}_{win} = \sum_{i \in s} \tilde{w}_i y_i^*, \tag{1}$$

gdzie zmodyfikowane wartości zmiennej badanej y_i^* są wyznaczane w następujący sposób:

$$y_i^* = \begin{cases} \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) K_{Ui}, & \text{jeśli } y_i > K_{Ui} \\ y_i, & \text{jeśli } K_{Ui} \geq y_i \geq K_{Li}, \\ \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) K_{Li}, & \text{jeśli } y_i < K_{Li}, \end{cases} \tag{2}$$

gdzie: $\tilde{w}_i = w_i g_i$

w_i – wagi zależne od schematu losowania, g_i – wagi zależne od próby,

$$g_i = 1 + \left(X - \hat{X}_{HT}\right)' \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}'_i\right)^{-1} (\mathbf{x}_i),$$

K_{Ui} – górny punkt graniczny, K_{Li} – dolny punkt graniczny,

$X = \sum_{i \in N} x_i$; $\hat{X}_{HT} = \sum_{i \in s} w_i x_i$ estymator bezpośredni HT wyznaczony dla zmiennej x.

Punkty graniczne wyznaczane są tak, by zminimalizować wartość MSE [Clark 1995]:

$$K_{Ui} = \mu_i^* - \frac{B_U}{(\tilde{w}_i - 1)}, \tag{3}$$

$$K_{Li} = \mu_i^* - \frac{B_L}{(\tilde{w}_i - 1)}, \tag{4}$$

gdzie $\mu_i^* = E(Y_i^*)$

$B_U = E\left[\hat{t}_{winU} - t_y\right]$ – obciążenie estymatora $\hat{t}_{winU}$,

$B_L = E\left[\hat{t}_{winL} - t_y\right]$ – obciążenie estymatora $\hat{t}_{winL}$,

$\hat{t}_{winU}$ – estymator Winsora wartości globalnej, uwzględniający tylko górny punkt graniczny,

$\hat{t}_{winL}$ – estymator Winsora wartości globalnej, uwzględniający tylko dolny punkt graniczny.

Ponieważ wartość μ_i^* jest trudna do oszacowania, w praktyce przyjmuje się, że $\mu_i^* = \mu_i$ [Mackin, Preston 2002]. Wówczas wartości graniczne szacowane są na podstawie wzorów:

$$K_{Ui} = \mu_i - \frac{B_U}{(\tilde{w}_i - 1)} = \mu_i + \frac{U}{(\tilde{w}_i - 1)}, \text{ gdzie } U = -B_U, \tag{5}$$

$$K_{Li} = \mu_i - \frac{B_L}{(\tilde{w}_i - 1)} = \mu_i + \frac{L}{(\tilde{w}_i - 1)}, \text{ gdzie } L = -B_L. \tag{6}$$

W celu wyznaczenia wartości U i L wykorzystuje się funkcję $\psi_U\left(\hat{D}_{(k)}\right)$ [Kokic, Bell 1994].

$$\psi_U\left(\hat{D}_{(k)}\right) = (k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)}, \text{ gdzie } \hat{D}_i = (Y_i - \hat{\mu}_i)(\tilde{w}_i - 1) \tag{7}$$

stanowi szacunek wartości reszt ważonych $D_i = (Y_i - \mu_i)(\tilde{w}_i - 1)$, $\hat{\mu}_i$ zaś oszacowanie parametru μ_i na podstawie regresji odpornej. Przed wyznaczeniem wartości funkcji $\psi_U\left(\hat{D}_{(k)}\right)$ należy uporządkować $\hat{D}_{(k)}$ tak, by $\hat{D}_{(1)} \geq \hat{D}_{(2)} \geq \ldots \geq 0 \geq \ldots$

Optymalną wartość U otrzymalibyśmy, rozwiązując równanie postaci $\psi_U(\hat{U}) = 0$. W praktyce, ze względu na trudność w znalezieniu właściwego rozwiązania równania, parametr U szacuje się, wykorzystując do tego celu następujący wzór:

$$\hat{U} = \frac{1}{(k^* + 1)} \sum_{j=1}^{k^*} \hat{D}_{(j)}, \tag{8}$$

gdzie k^* jest ostatnią wartością k, dla której wartość funkcji $\psi_U(\hat{D}_{(k)})$ nie jest ujemna.

Wartość L wyznaczana jest podobnie na podstawie funkcji:

$$\psi_L(\hat{D}_{(k)}) = (k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)}, \tag{9}$$

gdzie: $\hat{D}_{(k)}$ uporządkowane są wzrastająco: $\hat{D}_{(1)} \le \hat{D}_{(2)} \le ... \le 0 \le ...$

Drugim etapem w wyznaczaniu punktów granicznych jest oszacowanie na podstawie regresji odpornej wartości parametru μ_i. W tym celu w badaniu wykorzystano cztery metody.

Trimmed least squares **(TLS)**
Metoda ta polega na minimalizacji funkcji

$$F = \sum_{i=e} (y_i - \beta^T x_i)^2.$$

Na podstawie modelu obliczane są wartości reszt. Następnie z próby usuwane są te jednostki, dla których otrzymano największe dodatnie i ujemne reszty[1]. Dla tak zredukowanej próby wyznaczane są nowe parametry modelu.

Trimmed least absolute value **(LAV)**
Metoda polega na minimalizowaniu funkcji

$$F = \sum_{i=e} |y_i - \beta^T x_i|.$$

Na podstawie modelu obliczane są reszty. Podobnie jak w metodzie pierwszej z próby usuwane są te jednostki, dla których otrzymano największe reszty. Dla zredukowanej próby wyznaczane są nowe parametry modelu. Technika LAV powinna stanowić bardziej odporny model regresji niż TLS, gdyż duże wartości reszt w mniejszym stopniu wpływają na parametry regresji.

[1] We wszystkich technikach regresji odpornej zastosowanych w badaniu odsetek jednostek usuniętych był stały i wynosił 5%.

Technika *sample splitting* (TSS)

Technika *sample splitting* wykorzystuje KMNK. Stosowana jest ona do danych, które uprzednio w sposób losowy zostały podzielone na dwie równe części. Reszty dla jednej połowy danych wyznaczane są na podstawie modelu otrzymanego dla połowy drugiej. Po ponownym połączeniu próby usuwana jest jednostka, dla której zanotowano największą resztę. Ten proces jest powtarzany do momentu, aż określony odsetek jednostek zostanie usunięty. Technika ta powinna się charakteryzować większą odpornością niż TLS, gdyż reszty wykorzystywane przy usuwaniu jednostek nie są obliczane na podstawie modelu regresji, który był dla nich wyznaczony.

***Least median of squares* (LMS)**

Technika ta, opisana przez Rousseeuwa i Leroya [1987], polega na minimalizowaniu mediany z kwadratów reszt wyznaczonych na podstawie wartości z próby.

Przypomina to metodę bootstrapową. Polega na losowaniu podprób i obliczaniu dla każdej z nich wartości mediany kwadratów reszt, a następnie na wyborze modelu regresji z najmniejszą wartością mediany.

3. Schemat badania

W celu oceny technik regresji odpornej wykorzystywanych do szacunku parametru μ przeprowadzono badania symulacyjne dotyczące mikroprzedsiębiorstw.

Wykorzystano trzy źródła informacji:

Badanie reprezentacyjne SP3 – które zawierało informacje o zmiennych badanych. Próba wylosowana do badania SP3 w 2001 r. liczyła ponad 114 000 jednostek (4%). Jednak ostatecznie informacje pozyskano jedynie od 44 807 podmiotów gospodarczych.

Baza Jednostek Statystycznych (BJS) – zawierająca informacje z Krajowego Rejestru Urzędowego Podmiotów Gospodarki Narodowej REGON. Ogólna liczba małych podmiotów w systemie REGON w 2001 r. wynosiła 2 855 497.

Zbiory danych z systemu podatkowego Ministerstwa Finansów (rejestr podatkowy) – stanowiło 907 580 zeznań podatkowych od osób fizycznych i prawnych. Rejestr podatkowy wykorzystano jako źródło cech dodatkowych.

Szacunku dokonano dla zmiennej y – suma wypłaconych w ciągu roku wynagrodzeń brutto. Jako zmienną pomocniczą (x) wykorzystano przychody. Przy doborze zmiennej dodatkowej kierowano się przede wszystkim stopniem skorelowania informacji z badania SP3 oraz rejestru podatkowego.

Estymację przeprowadzono w przekroju: województwo/rodzaj prowadzonej działalności gospodarczej (sekcja PKD). Wyróżniono 160 domen (16 województw x 10 sekcji PKD).

W badaniu uwzględniono dwa podejścia, z których jedno polegało na wyznaczeniu dwóch punktów granicznych – górnego i dolnego, drugie zaś na wyznaczeniu jedynie górnego punktu granicznego.

Tabela 1. Wielkość próby w badaniu SP3 w przekroju województw i sekcji PKD

Województwo	A	B	C	D	E	F	G	H	I	J	K	M	N	O	Suma
Dolnośląskie	14	6	18	667	14	224	1069	45	174	381	292	61	129	96	3190
Kujawsko-pomorskie	16	4	14	321	14	105	635	36	109	20	119	15	82	56	1546
Lubelskie	32	6	16	363	14	132	683	39	143	453	126	31	128	78	2244
Lubuskie	13	9	9	306	9	86	479	31	104	46	127	12	106	61	1398
Łódzkie	17	4	20	729	17	243	921	49	153	262	186	65	134	175	2975
Małopolskie	33	3	15	771	11	414	1232	81	255	523	401	66	187	134	4126
Mazowieckie	18	8	25	975	16	575	2196	41	244	258	1420	325	211	614	6926
Opolskie	20	3	9	327	12	94	539	33	102	153	112	23	77	55	1559
Podkarpackie	14	6	17	398	20	107	1551	43	157	105	152	25	103	57	2755
Podlaskie	22	9	20	301	10	96	557	38	93	106	120	26	83	45	1526
Pomorskie	16	42	14	790	12	588	701	63	193	84	263	92	136	77	3071
Śląskie	25	7	12	1063	45	238	1982	91	311	385	315	87	305	368	5234
Świętokrzyskie	18	3	12	278	14	95	638	41	94	16	99	17	73	51	1449
Warmińsko-mazurskie	22	10	9	327	18	86	523	48	93	24	125	25	119	39	1468
Wielkopolskie	23	13	18	720	18	184	1118	61	205	287	382	59	166	92	3346
Zachodniopomorskie	13	34	13	338	14	176	611	63	170	119	200	26	144	73	1994
Suma całkowita	316	167	241	8674	258	3443	15 435	803	2600	3222	4439	955	2183	2071	44 807

A – leśnictwo, B – rybołówstwo i rybactwo, C – górnictwo i kopalnictwo, D – przetwórstwo przemysłowe, E – wytwarzanie i zaopatrzenie w energię, F – budownictwo, G – handel i naprawy, H – hotele i restauracje, I – transport, łączność, J – pośrednictwo finansowe, K – obsługa nieruchomości i firm, nauka, M – edukacja, N – ochrona zdrowia i opieka społeczna, O – pozostała działalność usługowa.

Źródło: obliczenia własne podstawie danych GUS z badania SP3.

4. Wyniki estymacji

Do wyznaczenia ocen precyzji badanych estymatorów zastosowano metodę bootstrapową. Wykonano 500 iteracji, na podstawie których wyznaczono wartość wariancji oraz współczynnika zmienności estymatora:

$$Var = \frac{1}{500-1}\sum_{b=1}^{500}\left(\hat{Y}_i - \hat{Y}\right)^2, \tag{10}$$

$$CV = \frac{\sqrt{Var}}{\hat{Y}}. \tag{11}$$

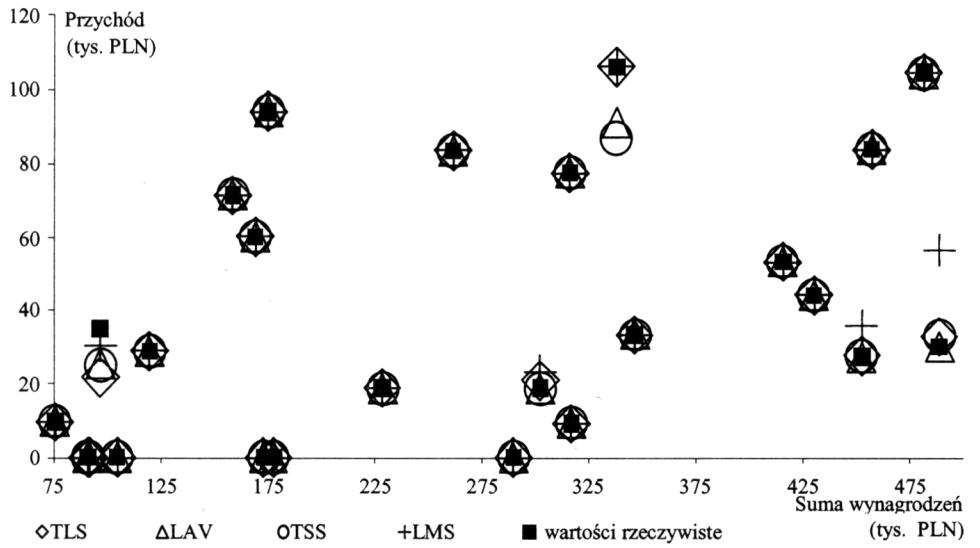
W ocenie rezultatów badania jako punkt odniesienia potraktowano klasyczną estymację bezpośrednią. Wyniki przeprowadzonego badania wskazują, że zasto-

sowanie technik regresji odpornej wpłynęło na redukcję współczynnika zmienności estymatora.

Tabela 2. Porównanie efektywności (CV) estymacji bezpośredniej z technikami regresji odpornej TLS, LAV, TSS, LMS

Regresja odporna	$\bar{C}V_i / \bar{C}V_{DIR}$ (%)	$\bar{C}V_i / \bar{C}V_{DIR}$ (%)
	dwa punkty graniczne	jeden punkt graniczny
TLS	0,804	0,893
LAV	0,798	0,887
TSS	**0,386**	**0,475**
LMS	0,843	0,960

Źródło: obliczenia własne.



Rys. 1. Diagram rozrzutu rzeczywistych wartości zmiennej badanej (przychodu) oraz wartości zmodyfikowanych na podstawie czterech technik regresji odpornej (TLS, LAV, TSS, LMS)
Źródło: opracowanie własne.

Zysk na efektywności był większy w przypadku wyznaczania dwóch punktów granicznych niż wówczas, gdy określano tylko górny punkt. Porównując wartości współczynników zmienności otrzymanych dla poszczególnych technik regresji, można zauważyć, że największa redukcja nastąpiła w przypadku techniki TSS (o ok. 60% dla 2 punktów i ok. 50% dla jednego punktu granicznego). Wartości CV dla dwóch technik (TLS i LAV) są zbliżone i w porównaniu z estymacją bezpośrednią niższe o ok. 20% dla 2 punktów oraz 10% dla jednego punktu. Największą zmiennością, a tym samym najmniejszą efektywnością charakteryzował się estymator, w którym wykorzystano techni-

kę LMS. W podejściu uwzględniającym jeden punkt graniczny wartość współczynnika zmienności była bardzo zbliżona do wyników estymacji bezpośredniej.

Diagram rozrzutu prezentuje rzeczywiste wartości zmiennej badanej oraz wartości zmodyfikowane na podstawie czterech zastosowanych w estymacji technik regresji odpornej (rys. 1).

Na wykresie widoczny jest zarówno kierunek, jak i zakres modyfikacji wartości zmiennej badanej. W przypadku jednostek uznanych za obserwacje odstające dla małych wartości sumy wypłacanych wynagrodzeń modyfikacja polegała na zmniejszeniu wartości zmiennej badanej. Zwiększenie wartości zmiennej badanej dotyczyło obserwacji odstających związanych z dużymi sumami wypłacanych wynagrodzeń. Zamianie wartości rzeczywistych zmiennej badanej na zmodyfikowane podlegał jedynie niewielki odsetek jednostek wylosowanych do próby (ok. 5%).

5. Wnioski

W artykule przedstawiono rezultaty badania symulacyjnego, w którym ocenie poddano różne techniki stosowane do wyznaczania wartości granicznych. W obliczeniach wykorzystano wyniki badania reprezentacyjnego SP3, dotyczącego mikroprzedsiębiorstw, wzbogacone o informacje pochodzące z rejestrów podatkowych Ministerstwa Finansów.

Estymacja Winsora dokonuje podziału jednostek wylosowanych do próby na dwie grupy: obserwacje odstające i pozostałe jednostki. W zależności od rodzaju badanej cechy wyznaczane są dwa punkty (górny i dolny) bądź jeden (górny) punkt graniczny. W przeprowadzonym badaniu przeanalizowano oba podejścia. Otrzymane wyniki wskazują, że w przypadku analizowanych zmiennych jakość szacunku jest wyższa, jeśli określi się dwa punkty graniczne.

Efektywność estymatora w przypadku estymacji Winsora zależy od wyboru punktów granicznych, a tym samym od rodzaju techniki regresji odpornej wykorzystanej do podziału próby na dwie grupy

Przeprowadzone badanie symulacyjne wykazało zależność między efektywnością szacunków a stopniem odporności techniki regresji. Zastosowanie techniki regresji bardziej odpornej wpływało na poprawę efektywności estymatora.

Literatura

Chambers R. (1996), *Robust Case-weighting for Multipurpose Establishment Surveys*, „Journal of Official Statistics", 12, s. 3-32.

Chambers R., Kokic P., Smith P., Cruddas M. (2000), *Winsorization for Identifying and Treating Outliers in Business Surveys*, Proceedings of the Second International Conference on Establishment Surveys (ICES II), s. 687-696.

Clark R.G. (1995), *Winsorization Methods in Sample Surveys*, Masters thesis, Department of Statistics, Australian National University.

Gross W.F., Bode G., Taylor J.M., Lloyd-Smith C.W. (1986), *Some Finite Population Estimators which Reduce the Contribution of Outliers*, Proceedings of the Pacific Statistical Conference, 20--24 May 1985, Auckland, New Zealand.

Hedlin D. (2004), *Business Survey Estimation,* R&D Report 2004:1, Statistics Sweden.

Hidiroglou M.H., Srinath K.P. (1981), *Some Estimators of Population Total from Simple Random Samples Containing Large Units*, „JASA" 76, s. 690-695.

Kokic P.N., Bell P.A. (1994), *Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator*, „Journal of Official Statistics" 10, s. 419-435.

Mackin C., Preston J. (2002), *Winsorization for Generalised Regression Estimation*, Australian Bureau of Statistics.

Rousseeuw P.J., Leroy P.M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons.

Samdal C.-E., Swensson B., Wretman J.H. (1992), *Model Assisted Survey Sampling*, Springer-Verlag.

Searls D.T. (1966), *An Estimator which Reduces Large True Observations*, „JASA" 61, s. 1200-1204.

WINSORIZATION IN SMALL BUSINESS SURVEY

Summary

Business data are often highly skewed to the right for two reasons: occurrence of outliers and a large proportion of zeroes. If by chance several unusually large residuals should fall in the sample then applying estimator may grossly underestimate or overestimate the population totals. One technique to deal with this problem is to divide a sample into two parts basing on cutoff values. Observations outside preset cutoff values are modified to values closer to these coutoff values. This estimator is called the winsorized estimator. The affectivity of winsorized estimator depends on the choice of the cutoff values, and hence the methods used to estimate regression parameters used to calculate these cutoff values. In this paper we examine the problem of the choice of one of the robust regression techniques to determine which techniques resulted in the best performing winsorized estimator. Simulation study presented here shows that *Sample Splitting Technique* results in the largest percentage reduction in MSE.