METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XV/4, 2014, str. 111 –120

WYBRANE TESTY STATYSTYCZNE DLA WARTOŚCI NIETYPOWYCH I ICH ZASTOSOWANIE W ANALIZACH EKONOMETRYCZNYCH

Dorota Pekasiewicz
Katedra Metod Statystycznych Uniwersytet Łódzki
e-mail: pekasiewicz@uni.lodz.pl

Streszczenie: Wykrywanie obserwacji nietypowych w próbie losowej stanowi ważne zagadnienie w analizach statystycznych. Jednym ze sposobów badania próby od kątem istnienia wartości odstających jest stosowanie testów statystycznych opartych na statystykach ekstremalnych, do których należą: test Grubbsa i jego uogólnienie, test Dixona oraz testy oparte na asymptotycznych rozkładach minimum i maksimum z próby. Granicznymi rozkładami statystyk ekstremalnych są, w zależności od klasy rozkładu analizowanej zmiennej, rozkład Gumbela, Frecheta lub Weibulla. W artykule, oprócz rozważań teoretycznych, przedstawiono zastosowania wybranych testów do weryfikacji hipotez o wartościach nietypowych przy konstrukcji modeli ekonometrycznych.

Słowa kluczowe: statystyki ekstremalne, minimum, maksimum, wartość nietypowa, rozkład Gumbela

WSTĘP

W konstrukcji modeli ekonometrycznych istotne znaczenie ma wykrywanie i ewentualnie eliminacja wartości nietypowych. W modelach regresji z jedną zmienną objaśniającą najprostszym sposobem ich wyznaczania jest analiza rozrzutu obserwacji (diagramu korelacyjnego), natomiast w przypadku regresji wielu zmiennych wykresów korelacyjnych par wszystkich zmiennych. Miernikami nietypowości są również reszty regresji oraz standaryzowane reszty regresji.

Ważną grupę metod służących wykrywaniu obserwacji odstających stanowią testy statystyczne. Wykorzystują one pojęcia statystyk ekstremalnych i ich własności, w szczególności twierdzenia o granicznych rozkładach minimum i maksimum z próby. Do testów weryfikujących hipotezy o istnieniu wartości

odstających wykorzystujących wartości ekstremalne należą m.in. test Grubbsa i jego uogólnienie oraz test Dixona. Informacja o klasie rozkładu statystyki minimum, czy maksimum rozważanej zmiennej losowej może być również wykorzystana przy konstrukcji statystyki testowej. Dla rozkładów o tzw. cienkich ogonach, takich jak rozkład normalny, logarytmiczno-normalny, granicznymi rozkładami statystyk ekstremalnych są rozkład Gumbela lub odwrócony Gumbela, natomiast dla rozkładów o grubych ogonach, np. *t*-Studenta, granicznymi rozkładami maksimum i minimum są, odpowiednio, rozkład Frecheta i odwrócony Frecheta. Dla zmiennych losowych przyjmujących wartości ze skończonego przedziału liczb rzeczywistych rozkłady statystyk ekstremalnych są typu Weibulla.

TEST GRUBBSA I JEGO UOGÓLNIENIE

Niech $X_1, X_2, ..., X_n$ będzie próbą losową pochodzącą z populacji o rozkładzie normalnym, zaś $x_1, x_2, ..., x_n$ wartościami wylosowanej próby.

Rozważmy następującą hipotezę:

H_0: w ciągu obserwacji $x_1, x_2, ..., x_n$ nie istnieją obserwacje nietypowe (o zbyt dużych lub zbyt małych wartościach).

Hipotezą alternatywną może być jedna z następujących hipotez:

H_1': minimalna wartość realizacji $x_1, x_2, ..., x_n$ próby losowej jest wartością nietypową,

H_1'': maksymalna wartość realizacji $x_1, x_2, ..., x_n$ próby losowej jest wartością nietypową.

Dla hipotez H_0 i H_1' statystyka testu Grubbsa [Grubbs 1969] wyraża się wzorem:

$$G_m = \frac{\overline{X} - X_{(1)}^{(n)}}{S}, \tag{1}$$

natomiast dla hipotez H_0 i H_1'' ma następującą postać:

$$G_M = \frac{X_{(n)}^{(n)} - \overline{X}}{S}, \tag{2}$$

gdzie $X_{(1)}^{(n)} = \min\{X_1, ..., X_n\}$ oraz $X_{(n)}^{(n)} = \max\{X_1, ..., X_n\}$.

Obszar krytyczny dla testu Grubbsa, przy przyjętym poziomie istotności α, wyznaczony jest przez wartość krytyczną:

$$G_\alpha = \frac{n-1}{\sqrt{n}} \frac{t_{(\alpha/n,n-2)}}{\sqrt{n-2+t_{(\alpha/n,n-2)}^2}}, \tag{3}$$

gdzie $t_{(\alpha/n,n-2)}$ jest wartością krytyczną rozkładu *t*-Studenta o $n-2$ stopniach swobody dla $\frac{\alpha}{n}$.

Hipotezę zerową odrzucamy na poziomie istotności α, gdy $G_M > G_\alpha$ lub $G_m > G_\alpha$.

Test Grubbsa rozstrzyga, czy wartość najmniejsza w próbie losowej lub wartość największa może być uznana za nietypową, odstającą od pozostałych elementów próby. Istnieje uogólnienie tego testu pozwalające wyznaczyć *r*-1 wartości nietypowych w *r* krokach testowania.

Procedurę weryfikacji hipotezy H_0 wobec $\sim H_0$ za pomocą uogólnionego testu Grubbsa przeprowadza się następująco.

W *i*-tym kroku, $i = 1, 2, ...,$ oblicza się:

- $\overline{X}_i = \frac{1}{n-i+1}\sum_{j\in L_i} X_j$, gdzie L_i jest zbiorem indeksów elementów próby zredukowanej, czyli powstałej poprzez pominięcie *i*-1 wartości,
- S_i - odchylenie standardowe wyznaczone w oparciu o zredukowaną próbę,
- statystykę testu:

$$G_i = \max_j \left(\frac{\left| X_j - \overline{X}_i \right|}{S_i} \right). \tag{4}$$

Statystykę tę porównuje się z wartością krytyczną:

$$G_{\alpha i} = \frac{n-i}{\sqrt{n-i+1}} \frac{t_{(p,n-i-1)}}{\sqrt{n-i-1+t^2_{(p,n-i-1)}}}, \tag{5}$$

gdzie $t_{(p,n-i-1)}$ jest wartością odczytaną z tablic wartości krytycznych rozkładu *t*-Studenta o n -*i* - 1 stopniach swobody dla wartości $p = \frac{\alpha}{n-i+1}$.

Spełnienie nierówności $G_i > G_{\alpha i}$ świadczy o istnieniu wartości nietypowej, którą pomijamy w kolejnym kroku. Testowanie hipotez kontynuujemy, aż do momentu otrzymania decyzji o braku podstaw do odrzucenia hipotezy zerowej. Jeżeli nastąpiło to w *r* krokach, to wyznaczonych zostało *r*-1 obserwacji nietypowych.

TEST DIXONA

Test Dixona jest kolejnym testem wykorzystującym wartości ekstremalne i weryfikującym hipotezy o istnieniu w próbie losowej $X_1, X_2, ..., X_n$ obserwacji nietypowych, przy założeniu, że pochodzi ona z populacji o rozkładzie normalnym [Dixon 1950, Chromiński, Tkacz 2010].

Dla hipotezy alternatywnej H_1' statystyka testu ma postać:

$$Q = \frac{X_{(2)}^{(n)} - X_{(1)}^{(n)}}{X_{(n)}^{(n)} - X_{(1)}^{(n)}}, \tag{6}$$

natomiast dla hipotezy alternatywnej H_1'':

$$Q = \frac{X_{(n)}^{(n)} - X_{(n-1)}^{(n)}}{X_{(n)}^{(n)} - X_{(1)}^{(n)}}, \tag{7}$$

gdzie $X_{(i)}^{(n)}$ oznacza *i*-tą statystykę pozycyjną (i = 1, 2, n-1, n).

Wartości krytyczne dla testu Dixona są stablicowane m.in. w pracy Verma, Quiroz-Ruiz [2006] dla $n \leq 100$.

TESTY OPARTE NA ASYMPTOTYCZNYCH ROZKŁADACH STATYSTYK EKSTREMALNYCH

Do konstrukcji testów weryfikujących hipotezy o istnieniu wartości nietypowych można wykorzystać informacje o granicznych rozkładach statystyk maksimum i minimum.

Niech X będzie populacją o rozkładzie normalnym $N(\mu, \sigma)$ oraz $X_1, X_2, ..., X_n$ próbą losową pochodzącą z tej populacji. Ponadto, niech:

$$Z_n = \max_{i \leq n}\left(\frac{X_i - \mu}{\sigma}\right), \tag{8}$$

$$W_n = \min_{i \leq n}\left(\frac{X_i - \mu}{\sigma}\right). \tag{9}$$

Standaryzowane zmienne Z_n i W_n mają, odpowiednio, rozkład Gumbela o dystrybuancie $H(z) = \exp(-\exp(-z))$ dla $z \in R$ i odwrócony rozkład Gumbela o dystrybuancie $L(z) = 1 - \exp(-\exp z)$, czyli

$$\lim_{n \to \infty} P\left\{\frac{Z_n - a_n}{b_n} < z\right\} = H(z), \tag{10}$$

oraz

$$\lim_{n\to\infty} P\left\{\frac{W_n - c_n}{b_n} < z\right\} = L(z), \tag{11}$$

gdzie $a_n = (2\ln n)^{1/2} - \frac{\ln\ln n + \ln 4\pi}{2(2\ln n)^{1/2}}$, $b_n = (2\ln n)^{-1/2}$ i $c_n = -a_n$ [Czekała 2001].

Zatem statystyka testu weryfikującego hipotezę H_0 wobec H_1'' wyraża się wzorem:

$$T_M = \frac{Z_n - a_n}{b_n}, \tag{12}$$

a wartość krytyczna wyznaczona z równania $\exp(-\exp(-z)) = 1-\alpha$ ma postać:

$$t_{M,\alpha} = -\ln\left(\ln\frac{1}{1-\alpha}\right). \tag{13}$$

Jeśli wartość t_M statystyki T_M spełnia nierówność $t_M \geq t_{M,\alpha}$, to na poziomie istotności α odrzucamy hipotezę H_0 na korzyść hipotezy alternatywnej mówiącej, że w próbie $X_1, X_2, ..., X_n$ istnieje obserwacja nietypowa o zbyt dużej wartości.

Dla hipotez H_0 i H_1' statystyka testu wyrażona jest wzorem:

$$T_m = \frac{\hat{W}_n + a_n}{b_n}. \tag{14}$$

Wartość t_m statystyki T_m porównujemy z wartością krytyczną $t_{m,\alpha}$:

$$t_{m,\alpha} = \ln\left(\ln\frac{1}{1-\alpha}\right). \tag{15}$$

Jeśli $t_m \geq t_{m,\alpha}$, to nie ma podstaw do odrzucenia hipotezy H_0, zaś jeśli $t_m < t_{m,\alpha}$, to odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej mówiącej, że istnieje wartość nietypowa w próbie o zbyt małej wartości.

Analogicznie skonstruować można testy weryfikujące hipotezy o istnieniu wartości odstających w próbie wylosowanej z populacji X o rozkładzie logarytmiczno-normalnym i logistycznym, gdyż odpowiednio wystandaryzowane statystyki maksimum i minimum mają również rozkład Gumbela lub odwrócony Gumbela. W przypadku rozkładu gamma, w szczególności wykładniczego oraz rozkładu Pareto statystyka maksimum ma rozkład Gumbela, a minimum Weibulla, natomiast dla rozkładu Cauchy'ego, t –Studenta odpowiednio standaryzowane statystyki ekstremalne mają rozkład Frecheta i odwrócony rozkład Frecheta [Castillo i in. 2004].

WYKRYWANIE I ELIMINACJA OBSERWACJI NIETYPOWYCH W PRZYPADKU MODELI EKONOMETRYCZNYCH

Rozważmy model ekonometryczny postaci:

$$Y_t = f(\mathbf{X_t}, \beta_0) + \varepsilon_t, \quad (t = 1, 2, ..., n) \tag{16}$$

gdzie: $(Y_t, \mathbf{X_t},)$ jest $(k+1)$ - elementowym wektorem losowym,

Y_t jest zmienną objaśnianą,

$\mathbf{X_t} = (X_{1t}, X_{2t}, ..., X_{kt})$ jest wektorem zmiennych objaśniających,

β_0 jest nieznanym wektorem parametrów strukturalnych,

ε_t jest składnikiem losowym, $\varepsilon_t \sim N(0, \sigma)$.

Analiza wielkości składnika losowego pozwala na wyznaczenie wartości nietypowych przy ustalaniu linii regresji i ewentualne ich pominięcie przy budowie modelu. Ze względu na często zakładaną normalność rozkładu zmiennej ε_t korzysta się z faktu, że standaryzowane statystyki maksimum i minimum składnika losowego mają rozkłady graniczne będące, odpowiednio, rozkładem Gumbela i odwróconym Gumbela.

Oszacowaniami statystyk $Z_n = \max_{t \le n}\left(\frac{\varepsilon_t}{\sigma}\right)$ oraz $W_n = \min_{t \le n}\left(\frac{\varepsilon_t}{\sigma}\right)$ na podstawie ciągu obserwacji $(y_1, x_{1,1}, x_{2,1}, ..., x_{k,1}), \ldots, (y_n, x_{1,n}, x_{2,n}, ..., x_{k,n})$ są statystyki $\hat{Z}_n = \max_{t \le n}\left(\frac{e_t}{\sigma}\right)$, $\hat{W}_n = \min_{t \le n}\left(\frac{e_t}{\sigma}\right)$, gdzie $e_t = y_t - f(\mathbf{x_t}, \beta_0)$.

W praktyce, w wielu przypadkach wariancja składnika losowego σ^2 jest nieznana i szacuje się ją za pomocą estymatora:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x_i}, \hat{\beta}_0)\right)^2. \tag{17}$$

Zastosowanie testów ze statystykami (12) i (14) sprowadza się do procedury sprawdzenia dla obserwacji (x_t, y_t) nierówności:

$$\frac{\frac{e_t}{s_n} - a_n}{b_n} > -\ln\left(\ln\frac{1}{1-\alpha}\right) \quad \text{lub} \quad \frac{\frac{e_t}{s_n} + a_n}{b_n} < \ln\left(\ln\frac{1}{1-\alpha}\right).$$

Problematykę wyznaczania wartości nietypowych można zilustrować na przykładzie konstrukcji modelu opisującego zależność dziennych stóp zwrotu

z inwestycji w akcje mBanku i PKO BP od stóp zwrotu z indeksu WIG 30 wprowadzonego na GPW w Warszawie dnia 23.09.2013 r. Rozważano następujący okres 23.09.2013 – 9.06.2014 r. Obserwacje z tego okresu stanowiły próbę 173 elementową.

Dla inwestycji w akcje mBanku otrzymano następującą funkcję (przy założeniu niezależności i normalności stóp zwrotu) :

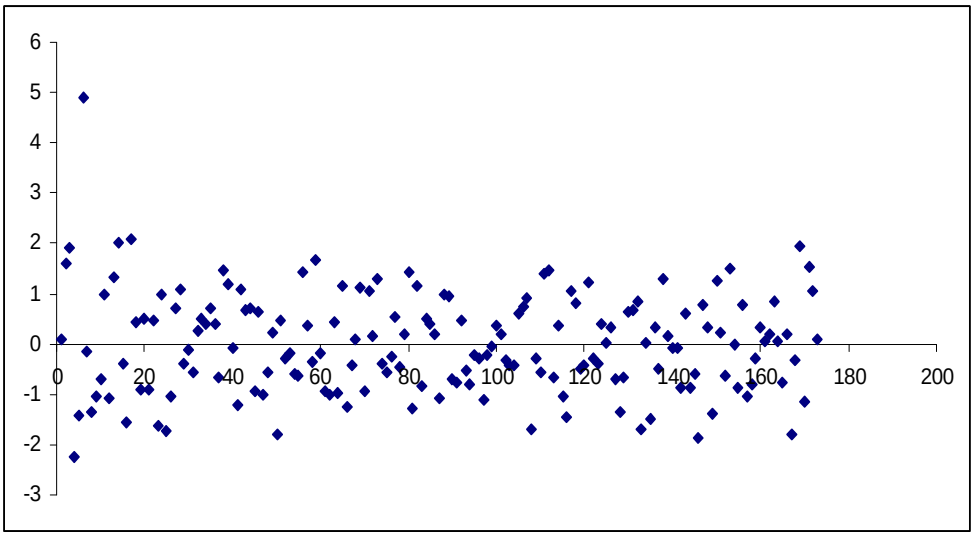
$$\hat{y}_t = 1{,}2966 x_t + 0{,}0009, \tag{18}$$

gdzie y_t - stopa zwrotu z inwestycji w akcje mBanku,

x_t - stopa zwrotu z indeksu WIG 30.

Zastosowanie testu opartego na granicznych rozkładach statystyk ekstremalnych związane jest z następującymi obliczeniami: $\hat{Z}_n = 4{,}8942$, $\hat{W}_n = -2{,}2279$, $a_n = 2{,}5608$, $b_n = 0{,}3115$. Statystyki (12) i (14) są równe: $t_M = 7{,}4909$, $t_m = 1{,}0690$, a wartości krytyczne wyznaczone dla poziomu istotności $\alpha = 0{,}05$: $t_{M,\alpha} = 2{,}9702$, $t_{m,\alpha} = -2{,}9702$. Ponieważ $t_M > t_{M,\alpha}$ parę wartości (0,0127; 0,0974), która pojawiła się 1.10.2013 r., należy uznać za nietypową przy konstrukcji modelu ekonometrycznego opisującego zależność stopy zwrotu z inwestycji w akcje mBanku od stopy zwrotu z indeksu giełdowego WIG 30 (*p- value*=0,0006). Na rysunku 1 przedstawiono standaryzowane reszty modelu opisującego zależność między rozważanymi wielkościami..

Rysunek 1. Standaryzowane reszty modelu (18)



Źródło: opracowanie własne

Ze względu na normalność rozkładu składnika losowego możliwe jest również zastosowanie testu Grubbsa. Związane jest to z obliczeniami: $e_{(1)}^{(n)} = \min\{e_{t1}, ..., e_{tn}\} = -0{,}0364$, $e_{(n)}^{(n)} = \max\{e_{t1}, ..., e_{tn}\} = 0{,}08$, $\bar{e}_t = 0$, $S_{e_t} = 0{,}0164$ i $G_M = 4{,}8947$, $G_m = 2{,}2273$ oraz $G_\alpha = 3{,}3878$ dla $\alpha = 0{,}05$. Na podstawie otrzymanych wyników wnioskuje się, że wartość 0,08 (standaryzowana reszta 4,8942) jest zbyt duża w porównaniu z pozostałymi co oznacza, że para wartości (0,0127; 0,0974) jest nietypowa, a to potwierdza decyzję podjętą na podstawie wcześniej stosowanego testu.

Analizując zależność wielkości stóp zwrotu z akcji PKO BP od stóp zwrotu z indeksu WIG 30 otrzymano następujący model:

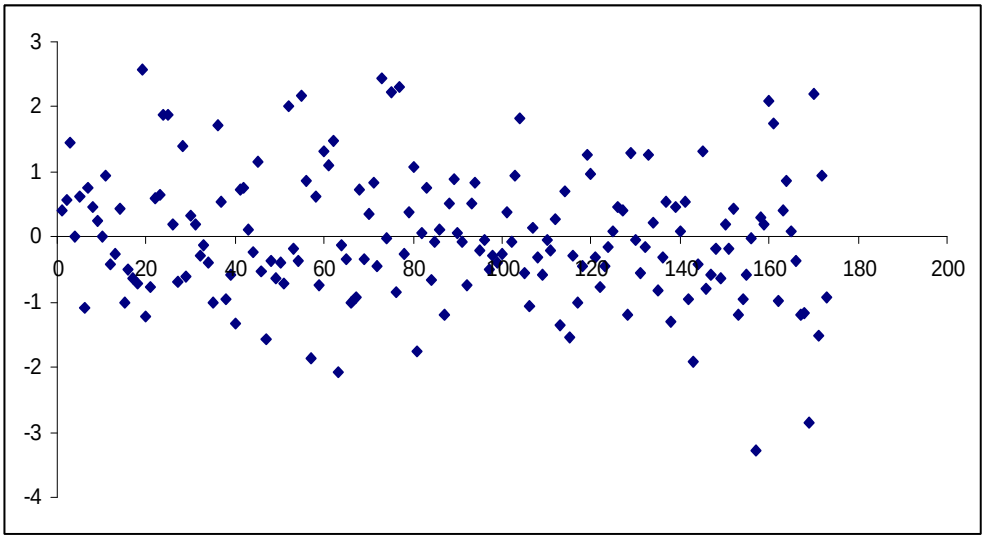
$$\hat{y}_t = 0{,}9143 x_t + 0{,}0005, \tag{19}$$

gdzie y_t - stopa zwrotu z inwestycji w akcje PKO BP,

x_t - stopa zwrotu z indeksu WIG 30.

W tym przypadku : $\hat{Z}_n = 2{,}5827$, $\hat{W}_n = -3{,}2780$ $a_n = 2{,}5608$, $b_n = 0{,}3115$. Stąd $t_M = 0{,}0703$, $t_m = -2{,}3022$ oraz $t_{M,\alpha} = 2{,}9702$, $t_{m,\alpha} = -2{,}9702$ dla $\alpha = 0{,}05$. Zatem na poziomie istotności 0,05 brak jest podstaw do odrzucenia hipotezy zerowej. Ponieważ *p- value*=0,0952, więc przyjmując $\alpha = 0{,}1$ parę wartości (0,0005; -0,0276) pochodzącą z dn. 16.05.2014 r. można uznać za nietypową ($t_{M,\alpha} = 2{,}2504$, $t_{m,\alpha} = -2{,}2504$). Graficzna prezentacja standaryzowanych reszt dla tego modelu przedstawiona jest na rysunku 2.

Rysunek 2. Standaryzowane reszty modelu (19)



Źródło: opracowanie własne

Przy zastosowaniu testu Grubbsa otrzymujemy: $e_{(1)}^{(n)} = \min\{e_{t1}, ..., e_{tn}\} = -0{,}0285$, $e_{(n)}^{(n)} = \max\{e_{t1}, ..., e_{tn}\} = 0{,}0224$, $\bar{e}_t = 0$, $S_{e_t} = 0{,}0087$ i $G_M = 2{,}5827$, $G_m = 3{,}2779$, $G_\alpha = 3{,}3878$ dla $\alpha = 0{,}05$ oraz $G_\alpha = 3{,}2045$ dla $\alpha = 0{,}1$. Stąd wnioskować można, na poziomie istotności 0,1, że wartość -0,0285 (standaryzowana reszta -3,278) jest zbyt mała, czyli para wartości (0,0005; -0,0276) jest nietypowa dla skonstruowanego modelu.

UWAGI KOŃCOWE

Wyznaczanie wartości nietypowych możliwe jest w analizach zmiennych losowych jednowymiarowych jak i wielowymiarowych, przy konstrukcji różnego rodzaju modeli ekonometrycznych. Zastosowanie testów Grubbsa, czy Dixona wiąże się z koniecznością sprawdzenia, czy zmienna losowa ma rozkład normalny. Testy wykorzystujące graniczne rozkłady statystyk ekstremalnych mają szersze zastosowanie. Znajomość rozkładu badanej zmiennej losowej, standaryzacja statystyk ekstremalnych i wykorzystanie twierdzeń mówiących o klasie ich rozkładów pozwalają konstruować tego typu testy.

W pracy przedstawiono testy statystyczne dla wartości odstających oparte na statystykach ekstremalnych. Inne testy np. Humpela, kwartylowy wykorzystujące statystyki pozycyjne takie jak mediana, czy kwartyle można znaleźć w pracy Budka i in. [2013].

Rozważane testy znajdują zastosowanie w badaniach zależności między dwiema lub większą liczbą zmiennych. Założenie o normalności rozkładu składnika losowego sprawia, że rozkładami granicznymi statystyk ekstremalnych są rozkład Gumbela i odwrócony Gumbela. Ich wykorzystanie, zarówno przy określaniu zależności wielkości stopy zwrotu akcji mBanku od stopy zwrotu indeksu WIG 30, jak i przy wyznaczaniu funkcji opisującej zależność wielkości stopy zwrotu akcji PKO BP od stopy zwrotu z indeksu WIG 30 prowadziło do uzyskania rezultatów identycznych z rezultatami testu Grubbsa. Porównywalne wyniki zachęcają do analiz testów opartych na granicznych rozkładach statystyk ekstremalnych dla zmiennych o innych rozkładach niż normalny, dla których zastosowanie testu Grubbsa, czy Dixona nie jest możliwe.

BIBLIOGRAFIA

Budka A., Kayzer D., Pietruczuk K., Szoszkiewicz K. [2013], Zastosowanie wybranych procedur do wykrywania obserwacji nietypowych w ocenie jakości rzek, PAN nr3/II, Warszawa.

Castillo E., Hadi A. S., Balakrishnan N., Sarabia J. M. [2004], Extreme value and related models with application in engineering and science, Wiley Interscience, A. John Wiley & Sons, Inc. New Jersey.

Chromiński K., Tkacz M. [2010], Comparison of outlier detection methods in biometric data, Journal of Medical Informatics and Technologies, 16, s. 89 – 94.

Czekała M. [2001], Statystyki pozycyjne w modelowaniu ekonometrycznym. Wybrane problemy Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Dixon W. J.[1950], Analysis of extreme values, Annals of Mathematical Statistics, 4, 488-506.

Grubbs F. [1969], Procedures for detecting outlying observations in samples, Technometrics 11, s. 1-21.

Verma S.P., Quiroz-Ruiz A. [2006], Critical values for six Dixon tests for outiers in normal samples up to sizes 100 and applications in science and engineering. Revista Mexicana de Ciencias Geologicas, 23(2), s.133 – 161.

CHOSEN STATISTICAL TESTS FOR OUTLIERS AND THEIR APPLICATION IN ECONOMETRIC ANALYSIS

Abstract: The problem of the existence of outliers in the sample is an important issue in statistical surveys. One of the methods of outliers detection is the application of statistical tests based on extreme statistics. Grubbs test and its generalization, Dixon test and tests based on asymptotic distributions of minimum and maximum (Gumbel, Frechet, Weibull distributions) belong to group of these tests. In the paper, besides the theoretical considerations the application of selected tests, used to verify the hypothesis of outliers in the construction of econometric models, is presented.

Keywords: extreme statistics, minimum, maximum, outlier, Gumbel distribution