Prace Naukowe Instytutu Matematyki i Informatyki Politechniki Częstochowskiej

ESTYMATORY WYKORZYSTUJĄCE INFORMACJĘ A PRIORI W ADAPTACYJNEJ ESTYMACJI PARAMETRU REGRESJI [1)]

Andrzej Grzybowski

Instytut Matematyki i Informatyki, Politechnika Częstochowska

Streszczenie: Omówiono problem adaptacyjnych procedur estymacji parametrów regresji. Zaproponowano pewne sekwencyjne estymatory, wykorzystujące informację a priori. W drugiej części pracy zbadano jakość proponowanych metod estymacji, stosując symulacje komputerowe.

Wstęp

W różnych współczesnych pracach naukowych poświęconych zarówno szeroko, jak i wąsko rozumianej teorii decyzji znajdziemy szereg wyrafinowanych reguł decyzyjnych, które w praktyce są bardzo rzadko wykorzystywane. Dzieje się tak między innymi dlatego, że dobre i pożądane własności tych reguł zależą od spełnienia rozmaitych, często subtelnych założeń tak o badanym zjawisku (np. postaci rozkładów losowych parametrów zjawiska), jak i o samym problemie w sensie teorii decyzji (np. od postaci funkcji strat czy klasy dostępnych reguł decyzyjnych). Często nie jest możliwe rozstrzygnięcie, który opis problemu jest najbardziej adekwatny w przypadku nas interesującym, a różne opisy prowadzą na ogół do zasadniczo różnych decyzji optymalnych. Dlatego tak naprawdę badacz wybiera regułę decyzyjną na podstawie własnych przeświadczeń o jej wartości i w oparciu o własną intuicję. Intuicja ta pochodzi głównie z doświadczeń praktycznych. Oddaje to często wygłaszana opinia, że wartość poszczególnych reguł decyzyjnych można ocenić dopiero w praktyce. Zapewne też z tych przyczyn pomimo istnienia wielu subtelnych i ważnych teoretycznie narzędzi służących do estymacji parametrów regresji, nadal zdecydowanie najczęściej używanym narzędziem jest, pochodząca jeszcze od Gaussa, metoda *minimum sumy kwadratów reszt*. Jest ona po prostu wielokrotnie sprawdzona w praktyce. Natura wielu zjawisk sprawia, że taki sposób weryfikacji *nowych* reguł decyzyjnych jest bardzo trudny i ryzykowny. Z jednej strony ewentualne eksperymentowanie może prowadzić do nadzwyczaj kosztownych konsekwencji, z drugiej strony w wielu dziedzinach zastosowań (np. ekonometria) okazja do eksperymentu nie zdarza się zbyt często. Wydaje się, że obecnie komputerowe badania symulacyjne stanowią ważną alternatywę jako sposób weryfikacji reguł decyzyjnych. Badania te mogą bardzo wzbogacić intuicję

[1)] Praca była realizowana w ramach projektu badawczego Nr 1 H02B 013 15 finansowanego przez Komitet Badań Naukowych.

badacza i pomóc w wyborze właściwej reguły. Dzięki nim badacz może nabrać osobistego przekonania, co do wartości poszczególnych reguł tak, jak gdyby sprawdzał je w praktyce. To osobiste zdanie na temat praktycznej wartości poszczególnych reguł ma zwykle znaczenie decydujące dla dokonywanego wyboru. W naszej pracy zajmiemy się pewnym szczególnym problemem analizy symulacyjnej reguł decyzyjnych. Dotyczy on predykcji wartości zmiennych zależnych na podstawie modeli regresji. Nie to jednak decyduje o szczególności rozpatrywanego problemu - zaproponowane podejście można stosować także do innych problemów, np. do sterowania układami. Decydują o owej szczególności dwa aspekty. Pierwszy to fakt, że symulacyjna ocena reguły decyzyjnej ma być dokonana na podstawie *zadanego, ograniczonego* zbioru danych. Drugi to fakt, że estymacja-predykcja ma być prowadzona *sekwencyjnie*.

2. Sformułowanie problemu

Rozważmy problem estymacji parametrów modelu zjawiska określonego równaniem

$$y = \beta^{\mathrm{T}}\mathbf{x} + Z$$

gdzie β oraz $\mathbf{x}$ to k-wymiarowe wektory parametrów odpowiednio regresji i zmiennych objaśniających. Zmienna losowa Z ma rozkład z wartością oczekiwaną 0 i ograniczonym drugim momentem.
Rozważmy sytuację, gdy zjawisko przez nas badane odbywa się regularnie w czasie i mamy możliwość obserwowania wartości zmiennych zależnej i niezależnych. W związku z tym mamy zbiór danych o rosnącej liczbie rekordów postaci (y_i, $x_{1i},...,x_{ki}$), gdzie i to numer rekordu. Niech wektor $\mathbf{Y}_n^m$ oraz macierz $\mathbf{X}_n^m$ oznaczają wartości zmiennych zależnej y_i i niezależnych $x_{1i},...,x_{ki}$ w momentach i od m do n. Gdy moment początkowy jest równy jeden ($m = 1$), wtedy indeks górny pomijamy. Często w sytuacjach rzeczywistych interesuje nas oszacowanie parametrów modelu regresji na kolejnych etapach realizacji obserwacji. Oznaczmy etapy (momenty) kolejnych estymacji jako e_1, e_2,...,e_n, a uzyskiwane na danym etapie oszacowania parametru β jako b1, b2,... . Załóżmy, że momenty e_1, e_2,...,e_n dokonywania kolejnych ocen parametru β są ustalone, zaś liczba etapów n dąży do nieskończoności. Przykładem tego typu problemów jest zagadnienie predykcji własności produktu na podstawie znanych parametrów procesu produkcyjnego.

Pierwszym problemem, który pojawia się w rozważanej sytuacji jest problem wykorzystania dotychczasowych obserwacji do estymacji parametru β. Przy niewielkiej liczbie obserwacji można by na każdym etapie e_i wykorzystywać zawsze całość obserwacji, tj. $\left(\mathrm{Y}_{e_i}, \mathbf{X}_{e_i}\right)$. W wielu rzeczywistych problemach nie jest to możliwe (liczba obserwacji rośnie nieograniczenie). Czy zatem należy część informacji odrzucić? Zdecydowanie nie, tym bardziej, że w praktyce oznaczałoby to

odrzucenie zdecydowanej większości obserwacji. Należy więc wskazać rodzaj estymacji sekwencyjnej, która będzie wykorzystywała wszystkie wyniki estymacji, ale pośrednio. Postać takich *adaptacyjnych* estymatorów wskażemy w następnym paragrafie. W dalszej części zaproponujemy również kryteria *symulacyjnej* oceny jakości wprowadzonych estymatorów w oparciu o *posiadane* rekordy, a więc na podstawie ograniczonego zbioru danych.

2. Estymatory adaptacyjne

Do sekwencyjnej estymacji parametrów regresji β na *i*-tym etapie proponujemy wykorzystać estymatory, których wartości są funkcją $\left(\mathbf{b}_{i-1}, \mathrm{Y}_{e_i}^{e_{i-1}}, \mathbf{X}_{\mathrm{e}_i}^{e_{i-1}}\right)$, czyli estymatory bazujące na oszacowaniu uzyskanym na etapie poprzednim oraz później uzyskanych obserwacjach. Estymator jest adaptacyjny w tym sensie, że jego wartość na kolejnym etapie zależy od wartości na etapie poprzednim - jest to więc zależność rekurencyjna. Oczywiście, jak w każdej zależności rekurencyjnej, konieczne jest określenie warunku początkowego. Podamy go za chwilę. Do estymacji adaptacyjnej parametrów regresji w opisanej sytuacji wykorzystamy estymatory postaci

$$\mathrm{d}_{(\Delta,\Sigma,\vartheta)}(\boldsymbol{Y}) = C(\Delta,\Sigma)\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{Y} + C(\Delta,\Sigma)\Delta^{-1}\vartheta \tag{2}$$

gdzie $C(\Delta,\Sigma) = (\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{X} + \Delta^{-1})^{-1}$. Estymatory tej postaci pojawiają się jako rozwiązania pewnych problemów estymacji bayesowskiej, a także pewnych problemów estymacji odpornej i minimaksowej, np. prace [7, 16, 19]. Wykorzystują one informację a priori o modelu regresji

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}$$

która ma postać (ϑ, Δ, Σ), gdzie wektor ϑ reprezentuje naszą informację a priori na temat parametru β, zaś macierz Δ odzwierciedla *niepewność* tej informacji. Dla przykładu, w analizie bayesowskiej ϑ byłoby wartością oczekiwaną rozkładu a priori parametru β, zaś Δ jego macierzą kowariancji. Macierz Σ reprezentuje naszą wiedzę o macierzy kowariancji zakłócenia **Z**. W serii prac [8-12] zajmowaliśmy się symulacyjną analizą sposobów wprowadzania informacji a priori do analizy regresji. Szczegółowo analizowaliśmy problemy, w których informacja a priori jest rezultatem wcześniejszej analizy regresji przeprowadzonej dla modelu takiego samego lub podobnego do opisującego badane przez nas zjawisko. Niech *bW* i *bA* będą oszacowaniami parametrów regresji β_W, β_A, otrzymanymi we *wcześniejszych* i *aktualnych* badaniach. Symbole s_W^2 i s_A^2 niech oznaczają *wcześniejsze* i *aktualne* oszacowanie wariancji zakłócenia *Z*. Okazuje się, że estymatory o dobrych własnościach otrzymujemy, gdy parametry ϑ i Σ - wzór (2) - są określone w następujący, intuicyjny sposób: $\Sigma = s_A^2\mathbf{I}_n$, (tzn. oszacowaniem na podstawie bieżących obserwacji), zaś wektor $\vartheta = bW$ (tzn. jest on określony zgodnie z naszą informacją

a priori). Mniej oczywista jest kwestia określenia parametru Δ mającego odzwierciedlać naszą niepewność związaną z posiadaną wiedzą na temat parametru. Z badań przedstawionych w pracach [6, 7, 9, 10] wynika, że jeśli macierz $\Delta = \Delta^* = [\delta^*_{ij}]_{k\times k}$ jest określona jako diagonalna z elementami $\delta^*_{ii} = bW_i^2$, to w przypadkach *rzetelnej* informacji a priori uzyskujemy bardzo dobre wyniki estymacji. Termin rzetelna informacja odnosi się do pewnych miar przydatności informacji a priori wprowadzonych w wymienionych wcześniej pracach. Można stwierdzić, że w tak rozumianym sensie informacja z poprzednich etapów estymacji w problemie przez nas rozważanym jest rzetelną informacją a priori dla etapów późniejszych. Zatem proponujemy jako estymator adaptacyjny na i-tym etapie estymator $d_{(\Delta,\Sigma,\vartheta)}(\mathbf{Y}_{e_i}^{e_{i-1}})$, w którym parametry Δ i ϑ zostały określone na bazie $\mathbf{b}_{e_{i-1}}$, zaś macierz Σ jest określona na podstawie $\mathbf{Y}_{e_i}^{e_{i-1}}$, $\mathbf{X}_{e_i}^{e_{i-1}}$. Jakość tak określonych estymatorów adaptacyjnych zostanie zweryfikowana symulacyjnie. Dla skrócenia dalszych zapisów estymator ten będziemy oznaczać $\mathbf{d}^*(i)$. Warunek początkowy tej rekurencji określamy następująco: na etapie pierwszym, zakładamy, że nie posiadamy informacji *a priori*, wykorzystamy estymator Gaussa-Markowa d_{LS} bazujący na pierwszych obserwacjach $\mathbf{Y}_{e_1}$, $\mathbf{X}_{e_1}$.

3. Badania krzyżowe - metodologia symulacyjnej analizy porównawczej

Analiza porównawcza *nowej* reguły decyzyjnej polega - oczywiście - na porównaniu wyników uzyskiwanych za pomocą badanej reguły decyzyjnej z wynikami otrzymanymi z wykorzystaniem innej reguły decyzyjnej. Ta *inna* reguła decyzyjna - oznaczmy ją $d0$ - stanowi punkt odniesienia. Reguła $d0$ powinna spełniać pewne warunki. Przede wszystkim powinna to być reguła klasyczna w tym sensie, że jest powszechnie znana, jej własności są dobrze uzasadnione teoretycznie i ma możliwie szerokie spektrum zastosowań. Metodologię symulacyjnej analizy porównawczej przeprowadzanej na podstawie zadanego, *ograniczonego* zbioru danych (obserwacji) określają:
- sposób wykorzystania posiadanego zbioru informacji,
- stosowane kryteria.

Jak to wielokrotnie podkreślaliśmy, zakładamy, że zbiór obserwacji, na podstawie którego chcemy dokonać *symulacyjnego* porównania metod, jest zbiorem ograniczonym. Rozumiemy to tak, że w momencie dokonywania porównania nie ma dodatkowych informacji i bazujemy na tych, które już mamy. Przypadek taki jest modelem wszystkich tych sytuacji, gdy analizujemy własności reguły na podstawie danych rzeczywistych, np. tak, jak u nas w dalszej części, pochodzących z produkcji danego wyrobu. Najczęściej wykorzystywaną wtedy techniką są badania krzyżowe. Spotykamy się z tą metodą w analizie regresji oraz w zastosowaniach sieci neuronowych. Można ją opisać w sposób następujący: Zbiór danych rzeczywistych - obserwacji - dzielimy na dwie części. Pierwsza z nich służy do

budowania modelu (u nas: estymacji parametru), druga zaś do sprawdzania jego jakości. Sprawdzanie polega na wykonaniu prognozy na podstawie obserwacji ze zbioru drugiego i porównaniu ich ze znanymi *prawdziwymi* realizacjami zmiennej zależnej. Jako miarę jakości oszacowania przyjmuje się średnią stratę uzyskaną we wszystkich przeprowadzonych prognozach. Pierwszy ze zbiorów będziemy nazywać zbiorem *obserwacji do estymacji* i oznaczać ZB_E (w literaturze angielskiej najczęściej spotykamy określenie *training set*). Drugi nazywamy zbiorem *obserwacji do weryfikacji jakości oszacowań* (po angielsku *validation set*). Oznaczać go będziemy symbolem ZB_W. Jeśli, dla ustalenia uwagi, przyjmiemy, że dostępny zbiór danych jest N-elementowy, a zbiór ZB_E ma N_E-elementów, to ocena jakości oszacowania jest średnią z $N_W = N\text{-}N_E$ prognoz weryfikowanych na podstawie zbioru ZB_W. Tak przeprowadzone badania krzyżowe noszą angielską nazwę *hold-out--set-cross-validation* (HOS-CV). Są one na ogół mało wiarygodne, gdyż uzyskana na ich podstawie ocena modelu może w znacznym stopniu zależeć od sposobu podziału zbioru danych na zbiory ZB_E i ZB_W, porównaj np. Kohavi [15], Plutowski i inni [17]. Zatem aby uczynić ją bardziej wiarygodną, powinniśmy proces ten powtórzyć dla większej liczby zbiorów ZB_E i jako ocenę ostateczną przyjąć średnią z ocen uzyskanych dla każdego z nich. W licznych pracach analizowane są teoretyczne i empiryczne własności różnych wariantów realizacji tej idei, tj. zarówno sposobu wyboru zbiorów, jak i ich liczby, porównaj np. Andrews [l], Breiman i Specter [5], Efron i Tibshirani [6], Plutowski i inni [17], Shao [18]. Ogólnie metody te określane są mianem *zostaw-wiele-do-weryfikacji.* Od angielskiej nazwy (*leave-many-out cross-validation*) oznaczać je będziemy akronimem LMO-CV. Do najważniejszych i najczęściej spotykanych ich wersji należą: *wyczerpująca* (*exhostive*) oraz *Monte Carlo* oznaczane, odpowiednio, ELMO-CV oraz MCLMO-CV.

W metodzie ELMO-CV dla ustalonej liczebności zbioru ZB_W przeprowadzamy HOS-CV dla każdego z nich. Jeśli jednak N nie jest wyjątkowo małe, to różnych zbiorów może być bardzo dużo, a przeprowadzenie obliczeń bardzo wyczerpujące. Tak np., jeśli $N = 50$ przyjmiemy $N_W = 10$, to różnych zbiorów ZB_W jest ponad 10 miliardów. Gdyby pojedynczy etap weryfikacji modelu (uzyskanie oszacowania i obliczenie średniego błędu prognozy dla jednego zbioru ZB_E) trwał l s, to pełna weryfikacja na podstawie wszystkich zbiorów 10-elementowych trwałaby, jak łatwo policzyć, ponad 325 lat! Dlatego w praktyce często przyjmuje się $N_W = l$. Jest to bardzo popularna wersja badań krzyżowych zwana *zostaw-jeden-do-weryfikcji* (*leave-one-out*). Oznaczać ją będziemy LOO-CV. Ma ona wiele zalet praktycznych, wśród których możemy wymienić ustaloną liczbę zbiorów do weryfikacji (mamy jedynie N różnych zbiorów ZB_W) oraz fakt, że liczebność zbioru do estymacji jest największa z możliwych. Nie bez znaczenia jest też porównywalność uzyskanych rezultatów wynikająca z faktu, że wszystkie podziały zbioru danych są jednoznaczne. Drugą ważną techniką realizacji badań krzyżowych jest MCLMO-CV. W podejściu tym proponuje się *wylosowanie* statystycznie reprezentatywnej liczby różnych zbiorów ZB_W o liczebności większej od l. Za każdym razem obserwacje tworzące ten zbiór losowane są bez zwracania. Jest to metoda zalecana przez wielu autorów, porównaj Breiman i Spector [l] Plutowski i inni [77], Shao [18]. W przy-

padku dużej liczby danych jest ona mniej kosztowna obliczeniowo od LOO-CV. Jest też bardziej odporna na błędy w danych. Porównanie własności tych i innych technik badań krzyżowych można znaleźć w pracach Breimana i Spectora [5], Efrona i Tibshiraniego [6], Kohaviego [15], Plutowskiego i innych [17], Shao [18]. Ostatnia z wymienionych prac zawiera bardzo ciekawe wyniki wskazujące na liczne dobre własności metody MCLMO-CV.
Ze względu na jej zalety w naszych dalszych przykładach będziemy stosować metodę MCLMO-CV.

4. Kryteria jakości

Ponieważ naszym celem jest porównanie własności zaproponowanych metod estymacji z innymi, należy wskazać *te inne*. W pracy proponujemy jako estymator odniesienia d_0 przyjęcie estymatora Gaussa-Markowa d_{LS}. Ze względu na zakres i częstość zastosowań praktycznych oraz stałe ryzyko z nim związane jest to najczęściej wykorzystywany w badaniach porównawczych estymator. Oczywiście w naszym przypadku zaproponujemy pewne sposoby adaptacyjnego wykorzystania tego estymatora, tak jak często jest to realizowane w praktyce. Ze sposobem wykorzystywania estymatora d_{LS} związane są poniższe kryteria jakości zaproponowanej przez nas metody estymacji.
Kryterium 1 (K1). Zbiór danych dzielony jest losowo na dwie części. Na podstawie części pierwszej dokonujemy estymacji parametrów za pomocą *n*-etapowej estymacji adaptacyjnej oraz za pomocą estymatora $\mathbf{d}_{\mathrm{LS}}$. Uzyskane oceny wykorzystywane są do prognozowania wartości zmiennej zależnej w drugiej części danych. *Wartością kryterium* jest wektor średnich strat obu predyktorów oraz wariancji tych strat.
Kryterium 2 (K2). Zbiór danych jest dzielony losowo na $n+1$ równych części - odpowiadających danym do n etapów estymacji adaptacyjnej. Na i-tym ($i = 1,...,n$) etapie:
- estymujemy parametry regresji za pomocą estymatora $\mathbf{d}^*$,
- estymujemy parametry regresji za pomocą estymatora d_{LS} na podstawie całości dotychczasowych danych, tj. $\mathbf{Y}_{e_i}$, $\mathbf{X}_{e_i}$.

Uzyskane oceny wykorzystujemy do prognozowania wartości zmiennej zależnej w części danych dla etapu następnego i rejestrowane są średnie straty obu predyktorów oraz inne charakterystyki statystyczne. *Wartość kryterium* to wektor średnich strat obu estymatorów w całym eksperymencie oraz wariancje tych strat.
Kryterium 3 (K3). Zbiór danych, podobnie jak w K2, jest dzielony losowo na $n+1$ możliwie równych części - odpowiadających danym do n etapów estymacji adaptacyjnej. Na i-tym etapie:
- estymujemy parametry regresji za pomocą estymatora $\mathbf{d}^*$,
- estymujemy parametry regresji za pomocą estymatora $\mathbf{d}_{\mathrm{LS}}$ na podstawie danych tego etapu dotychczasowych danych, tj. $\mathbf{Y}_{e_i}^{e_{i-1}}$, $\mathbf{X}_{e_i}^{e_{i-1}}$.

Uzyskane oceny wykorzystujemy do prognozowania wartości zmiennej zależnej w części danych dla etapu następnego i rejestrowane są średnie straty obu predyktorów oraz inne charakterystyki statystyczne. *Wartość kryterium* to wektor średnich strat obu estymatorów w całym eksperymencie oraz wariancje tych strat.

Zauważmy, że w kryteriach **K1** i **K2** porównuje się zaproponowany estymator adaptacyjny z estymatorem bazującym na całej dotychczasowej informacji. Jest to sytuacja idealna, która w praktyce nie może mieć miejsca, ze względu na wspomnianą wcześniej dążąca do nieskończoności liczbę obserwacji. Zatem w rzeczywistości nie jest to porównanie z metodą „konkurencyjną". Natomiast jako rzeczywiście konkurencyjną można uznać metodę zaproponowaną w kryterium trzecim.

Zauważmy również, że wartościami kryteriów są pewne parametry statystyczne. Nie stanowią one same w sobie wartości kryterialnej w sensie porządkującym czy wartościującym. Dopiero statystyczna analiza otrzymanych wartości może prowadzić do wybrania reguły decyzyjnej.

5. Symulacyjna analiza porównawcza

Przykłady przedstawione w tym paragrafie ilustrują zastosowanie wskazanej metodologii do oceny jakości zdefiniowanych w paragrafie 2 estymatorów adaptacyjnych. Dane służące do porównań w przykładach 1-3 były danymi sztucznymi, w przykładzie 4 analizujemy własności estymatora **d***, wykorzystując dane rzeczywiste. Zarówno program generowania danych, jak i wszystkie procedury porównujące zostały zaprogramowane w języku wewnętrznym pakietu *Mathematica* 4.0.

W przykładach 1-3 liczba rekordów wynosiła 1600, liczba etapów estymacji adaptacyjnej 15. Liczba rekordów do sprawdzania jakości uzyskanych ocen w sensie kryterium K1 wynosiła 100, a liczba generacji różnych podziałów na zbiory do oszacowań i do oceny jakości wynosiła 100. We wszystkich przykładach funkcja start to odległość (mierzona wartością bezwzględną) pomiędzy *prognozą* $\hat{a}$ a *wartością rzeczywistą a*, tj.

$$L(a,\hat{a}) = |a - \hat{a}|$$

We wszystkich przykładach jako miarę różnicy między dwoma efektami estymacji uzyskanymi w dwóch porównywanych podejściach przyjęliśmy wartość statystyki

$$T = \frac{\overline{L}_1 - \overline{L}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

gdzie $\overline{L}_i, S_i^2$ oznaczają średnią stratę, wariancję empiryczną średnich strat *i*-tej metody estymacji. Liczba n_i oznacza w przypadku kryterium K1 liczbę losowań podziałów zbioru danych, w przypadku zaś kryteriów K2 i K3 liczbę etapów estymacji pomnożoną przez liczbę losowań podziałów do tych etapów. Statystyka

analogiczna wykorzystywana jest w teście równości wartości średnich dwóch rozkładów. Oczywiście w naszych przykładach nie są (a w każdym razie nie muszą być) spełnione założenia konieczne do znajomości rozkładu tej statystyki. Można przypuszczać jednak, że ze względu na bardzo dużą liczbę obserwacji rozkład tej statystyki może być wystarczająco dokładnie dla naszych celów przybliżony rozkładem normalnym. Tak czy owak jest to miara obiektywna, dająca dość dokładne pojęcie o istotności różnic występujących w rezultatach uzyskanych za pomocą porównywanych metod.

Przykład 1

Dane analizowane w tym przykładzie były danymi generowanymi w modelu liniowym, z zakłóceniem losowym o odchyleniu standardowym nieprzekraczającym 10% średniego poziomu zmiennej zależnej. Były to więc dane „korzystne" dla estymatora d_{LS}. Uzyskano następujące wartości statystyki w poszczególnych kryteriach:

$$\text{K1: } T = 0.92 \qquad \text{K2: } T = 1.6 \qquad \text{K3: } T = -1.87$$

Przyjmując, na podstawie rozkładu normalnego, że istotne różnice występują dla wartości ponad 1,25, otrzymujemy wniosek, że w sytuacji gdyby można było korzystać z całości danych na każdym etapie, wtedy procedura oparta na estymatorze d_{LS} i przedstawiona w kryterium K2 jest najlepsza (porównajmy K1 i K2). W sytuacji gdy problem wymaga estymacji rzeczywiście adaptacyjnej, procedura zaproponowana w tej pracy ma w tym przypadku istotną przewagę.

Przykład 2

Dane analizowane w kolejnym przykładzie, były danymi generowanymi w modelu nieliniowym (zmienna zależna była funkcja potęgową dwóch zmiennych objaśniających z potęgami 3 i 4), z zakłóceniem losowym o odchyleniu standardowym nieprzekraczającym 10% średniego poziomu zmiennej zależnej. Uzyskano następujące wartości statystyki T w poszczególnych kryteriach:

$$\text{K1: } T = 0.78 \qquad \text{K2: } T = 2.49 \qquad \text{K3: } T = -1.29$$

Wnioski, jakie się nasuwają po porównaniu tych wskaźników, są analogiczne jak w przykładzie poprzednim. W sytuacji gdy problem wymaga estymacji istotnie adaptacyjnej, procedura zaproponowana w tej pracy ma i w tym przypadku istotną przewagę.

Przykład 3

Tym razem dane były generowane w modelu mniej nieliniowym jak w przykładzie drugim (potęgi były równe 3/2) z zakłóceniem losowym o odchyleniu standardowym nieprzekraczającym 10% średniego poziomu zmiennej zależnej. Różnica polegała na tym, że 5% obserwacji było generowanych absolutnie losowo - niezgodnie z modelem. Były to więc dane z zaburzeniami. Uzyskano następujące wartości statystyki w poszczególnych kryteriach:

$$\text{K1: } T = -1.16 \qquad \text{K2: } T = -2.6 \qquad \text{K3: } T = -2.73$$

Jak widzimy, tym razem nawet wykorzystanie pełnej informacji w estymatorze Gaussa-Markowa nie daje mu przewagi. Procedura adaptacyjna jest istotnie lepsza według każdego kryterium.

Przykład 4 - dane rzeczywiste

W tym przykładzie przeanalizujemy własności estymatora d* na podstawie danych rzeczywistych. Rozważać będziemy dane dotyczące własności blach otrzymywanych w wyniku procesu technologicznego realizowanego w hutach. Własności te są opisywane różnymi parametrami między innymi przez tzw. granicę wytrzymałości, granicę plastyczności, wydłużenie i udarność, porównaj np. prace Grzybowskiego i Urbanowicza [13, 14]. Tutaj skupimy się na modelu granicy wytrzymałości blachy *Re*. W drugiej z cytowanych prac zaproponowano następującą postać modelu zależności wskaźnika Re od wybranych parametrów chemicznych i technologicznych:

$$Re = \beta_1 + \beta_2 C + \beta_3\, Mn + \beta_4 Si + \beta_5 Ni + \beta_6 Al + \beta_7 Nb + \beta_8 Wf + \beta_9 Gr + \beta_{10} Tk + Z$$

gdzie *Gr*, *Wf*, *Tk* są pewnymi parametrami charakteryzującymi proces produkcyjny, zaś pozostałe oznaczenia to symbole chemiczne pierwiastków wchodzących w skład produktu.

Przeprowadzając analizę porównawczą dla 503 rekordów pochodzących z produkcji walcowni, otrzymaliśmy następujące wartości statystyki T:

K1: T = 4.4 K2: T = 5.2 K3: T = –6.0

W tym przypadku, podobnie jak w pierwszym przykładzie, otrzymujemy wniosek, że w sytuacji gdy można korzystać z całości danych na każdym etapie, procedura oparta na estymatorze d_{LS} i przedstawiona w kryterium K2 jest najlepsza. Jednak oczywiście w rozważanej tutaj sytuacji jest to niemożliwe - ilość danych rośnie nieograniczenie i to w szybkim tempie. Zatem problem wymaga estymacji istotnie adaptacyjnej, a wtedy, jak widzimy, procedura zaproponowana w tej pracy ma bardzo wyraźną przewagę.

Podsumowanie

Jak widzimy, przydatność proponowanej metody estymacji adaptacyjnej zależy od charakteru problemu rzeczywistego (i związanego z tym charakteru danych *generowanych* przez ten problem). Wskazane jest więc, by zaproponowaną metodologię badań symulacyjnych przeprowadzać na rzeczywistych danych w każdym konkretnym przypadku ewentualnych zastosowań wskazanych metod estymacji. Pozwoli to przekonać się, czy w danym przypadku warto i należy stosować proponowaną adaptacyjną metodę estymacji. Jak wskazują wyniki tutaj przedstawione, metoda analizowaną w tej pracy jest metodą konkurencyjną i w wielu sytuacjach rzeczywistych na pewno wartą implementacji.

Literatura

[1] Andrews D.W.K., Asymptotic optimality of generalized CL, cross-validation, and generalized cross--validation in regression with heteroskedastic errors, Journal of Econometrics 1991, 47, 359-377.

[2] Berger J., A robust generalised bayes estimator and confidence region for a multivariate normal mean, Ann. Statist. 1980, 8, 716-761.

[3] Berger J., Selecting a minimax estimator of a multivariate normal mean, Ann. Statist. 1982, 10, 81-92.

[4] Berger J., Robust Bayesian analysis: sensitivity to the prior, J. Statist. Plann. Inference 1990, 25, 303-328.

[5] Breiman L., Spector P., Submodel selection and evaluation in regression: the X-random case, Internation Statistical Review 1992, 60, 3, 291-319.

[6] Efron B., Tibshirani R.J., Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule, Technical Raport 176, Dept. of Statistics, Stanford Univ., 1995.

[7] Grzybowski A., Minimaksowo-odporna estymacja parametrów regresji, Przegląd Statystyczny 1997, R. XLIV, 3, 427-435.

[8] Grzybowski A., Symulacyjna metoda doboru parametrów niepewności w minimaksowo-odpornej analizie regresji, (w:) Metody i zastosowania badań operacyjnych, praca zbiorowa pod red. T. Trzaskalika, cz. II, Wydawnictwo Uczelniane Akademii Ekonomicznej w Katowicach, Katowice 1998, 267-278.

[9] Grzybowski A., Simulation analysis of some regression estimators incorporating prior information, Proceedings of International Workshop On Statistical Modelling, Graz 1999.

[10] Grzybowski A., Symulacje komputerowe w analizie metod estymacji parametrów regresji wykorzystujących informację a priori, Materiały 6 Warsztatów Naukowych Symulacja w Badaniach i Rozwoju, Białystok 25-27.08.1999.

[11] Grzybowski A., Comparative analysis of some nonlinear regression estimators incorporating prior information - simulation studies, Proceedings of 3rd International Conference on Parallel Processing & Applied Mathematics, Kazimierz Dolny 14-17.09.1999, 525-532.

[12] Grzybowski A., Simulation analysis of some regression estimators incorporating prior information - performance for different loss functions, Proceedings of 16th IMACS World Congress 2000 on Scientific Computation, Applied Mathematics and Simulation, Lausanne 2000.

[13] Grzybowski A., Urbanowicz Z., Statystyczne modelowanie własności mechanicznych blach grubych przy regulowanym walcowaniu, Materiały Konferencji Zastosowania Komputerów w Zakładach Przetwórstwa Metali, Bukowina 1998, 25-33.

[14] Grzybowski A., Urbanowicz Z., Alternative methods of regression in modelling properties of steel plates - a comparative studies, Proceedings of 3rd International Conference on Parallel Processing & Applied Mathematics, Kazimierz Dolny 1999, 533-542.

[15] Kohavi R., A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Mateo 1995, 1137-1143.

[16] Milo W., Stabilność i wrażliwość metod ekonometrycznych, Łódź 1995.

[17] Plutowski M.E., Sakata S., White H., Cross-validation estimates integrated mean squared error, Advances in neural information processing systems 6, eds. C.L. Giles, S.J. Hanson, J.D. Cowan, San Mateo, CA, Morgan Kaufmann Publishers 1994.

[18] Shao J., Linear model selection by cross-validation, Journal of the American Statistical Association 1993, 88, 422, 486-494.

[19] Stahlecker P.,Trankler G., (ed.)., Acta Applicandae Mathematicae 1996, 43, 1.

[20] Schipp B., Trankler G., Stahleker P., Minimax estimation with additional linear restrictions - a simulation study, Commun. Statist. - Simula. 1988, 17(2), 393-406.