PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Mariusz Kubus
Politechnika Opolska

ANALIZA METODY LARS W PROBLEMIE SELEKCJI ZMIENNYCH W REGRESJI

Streszczenie: Selekcja zmiennych jest typowym zadaniem *data mining*, gdzie prowadzący analizę poszukuje interesujących i nieoczekiwanych relacji w danych bez wiedzy początkowej na temat badanego zjawiska. W liniowym modelu regresji, zamiast popularnej procedury krokowej czy też eliminacji zmiennych testem istotności współczynników, do selekcji zmiennych zastosować można metody iteracyjnej estymacji parametrów modelu (np. LARS Efrona i in. [2004]). Celem artykułu jest zbadanie zdolności metody LARS do identyfikowania zmiennych nieistotnych, szczególnie gdy zachodzą między nimi zależności liniowe. Dokonano w nim też porównania z wybranymi metodami selekcji zmiennych.

Słowa kluczowe: selekcja zmiennych, selekcja w algorytmie uczącym, metoda LARS.

1. Wstęp

Problem selekcji zmiennych nabiera wagi szczególnie w sytuacjach, gdy celem prowadzącego analizę jest wydobycie z danych interesujących i nieoczekiwanych związków, bez wiedzy początkowej na temat badanego zjawiska. Wybór relatywnie niewielkiego zestawu najistotniejszych predyktorów ma nie tylko walory interpretacyjne, ale może też poprawić predykcję.

Obecnie w literaturze z zakresu selekcji zmiennych (np. [Guyon i in. 2006; Liu, Yu 2005; Guyon, Elisseeff 2003; Blum, Langley 1997]) przyjmuje się następującą klasyfikację podejść do tego zagadnienia. Podejście pierwsze nazywane filtrowaniem zmiennych polega na przeprowadzeniu selekcji zmiennych przed etapem budowania modelu. W regresji zwykle wykorzystuje się na różne sposoby współczynnik korelacji Pearsona. Podejście drugie (*wrappers*) polega na wykorzystaniu algorytmu uczącego do oceny różnych podzbiorów zmiennych. Zadanie można sformułować jako zagadnienie przeszukiwania przestrzeni wszystkich kombinacji cech. Do jego rozwiązania wybiera się technikę przeszukiwania heurystycznego (przeszukanie wszystkich kombinacji zmiennych przy dużym wymiarze nie wcho-

dzi w rachubę) oraz funkcję oceniającą podzbiory zmiennych. Typowym przykładem jest tu regresja krokowa, która wykorzystuje powszechnie w *data mining* stosowaną strategię zachłanną, dobierając w każdym kroku do aktualnego zestawu zmiennych jedną zmienną lub eliminując jedną zmienną. Przeszukiwanie jest sterowane funkcją kryterium; w jej charakterze można wykorzystać częściowy test F, kryterium informacyjne AIC lub posłużyć się błędem predykcji szacowanym przez sprawdzanie krzyżowe (lub na odrębnym zbiorze walidacyjnym). Odmienne jest trzecie podejście do selekcji zmiennych (*embedded methods*), polegające na zastosowaniu metod, w których selekcja jest integralną częścią (np. drzewa regresyjne, LASSO czy LARS).

Zaproponowana przez Efrona i in. [2004] metoda LARS polega na iteracyjnej estymacji współczynników regresji liniowej. Mieści się ona zarazem w nurcie modelowania addytywnego, gdzie w każdym kroku kolejna funkcja składowa poprawia dopasowanie modelu do danych poprzez zastosowanie regresji aktualnych reszt, a więc niewyjaśnionej jeszcze części zmienności zmiennej objaśnianej. W każdym kroku do modelu włączana jest kolejna zmienna objaśniająca. Algorytm wykonuje więc p kroków. W celu wyeliminowania zmiennych nieistotnych można wprowadzić kryterium stopu lub na podstawie wybranej funkcji oceny wybrać postać modelu (ze względu na liczbę zmiennych) na końcu.

Celem artykułu jest zbadanie odporności metody LARS na zmienne nieistotne, które będą generowane tak, by występowały między nimi związki liniowe, co wywołuje problemy w oszacowaniu parametrów klasycznego, liniowego modelu regresji wielorakiej.

2. LASSO i iteracyjne metody estymacji parametrów modelu liniowego

Selekcja zmiennych poprzez regresję krokową lub eliminacja zmiennych testem istotności współczynników może być postrzegana jako nadawanie zmiennym wag. Zmienne eliminowane otrzymują wagę 0, a pozostałe według metody najmniejszych kwadratów (MNK). Podejściem alternatywnym jest regresja regularyzowana, w szczególności LASSO (*Least Absolute Shrinkage and Selection Operator*) [Tibshirani 1996], gdzie część współczynników się zeruje, a część ma mniejszą (co do wartości bezwzględnej) wartość od odpowiadających im współczynników MNK.

Zagadnienie LASSO można sformułować:

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 , \text{ z ograniczeniem: } \sum_{j=1}^{p} \left| \beta_j \right| \le t . \quad (1)$$

Niestety, nie istnieje tu macierzowe rozwiązanie w zamkniętej postaci (tak jak w regresji grzbietowej). Do rozwiązania problemu stosuje się programowanie kwadratowe z liniowymi ograniczeniami lub metody przybliżone (zob. np. [Hastie i in. 2009; Miller 2002, Osborne i in. 2000]).

Podstawowym problemem praktycznego stosowania metody LASSO jest wybór właściwej wartości progowej *t* w ograniczeniu (1). Ustalenie *t* większego (lub równego) od sumy wartości bezwzględnych estymatorów metody najmniejszych kwadratów $\hat{\beta}_j^{MNK}$ prowadzi do rozwiązania identycznego z MNK. Dopiero zmniejszanie wartości *t* skutkuje zerowaniem się niektórych współczynników, co w efekcie prowadzi do uzyskania modeli z różną liczbą zmiennych objaśniających. W skrajnym przypadku, gdy $t = 0$, otrzymuje się model zbudowany tylko z wyrazu wolnego (przewidujący wartości zmiennej objaśnianej na podstawie jej średniej z próby uczącej). W celu wybrania właściwej wartości progowej *t* wybiera się *k* wartości pośrednich między 0 a $\sum_{j=1}^{p}\left|\hat{\beta}_j^{MNK}\right|$ i buduje *k* modeli. Ocena tych modeli decyduje o wyborze wartości progowej *t*. Z kolei do oceny modeli można wybrać miarę wykorzystującą kompromis między dokładnością a złożonością modelu (np. kryteria informacyjne) lub posłużyć się błędem predykcji estymowanym przez sprawdzanie krzyżowe (lub na odrębnym zbiorze walidacyjnym). Dodatkowo można zastosować regułę jednego błędu standardowego.

Niemal identyczne z LASSO rozwiązania można uzyskać za pomocą algorytmu modelowania addytywnego ze skracaniem (*forward stagewise with shrinkage*) [Hastie i in. 2009]. Główną jego ideą jest przeprowadzenie ciągu regresji prostych aktualnych reszt względem najbardziej skorelowanego predyktora. W ten sposób model dopasowywany jest w kolejnych krokach do niewyjaśnionej dotąd części zmienności zmiennej objaśnianej *y*.

Na początku zmienne objaśniające są standaryzowane, a wartości początkowe współczynników $\hat{\beta}_1 = \hat{\beta}_2 = ... = \hat{\beta}_p = 0$. Wyraz wolny jest estymowany przez średnią z próby uczącej $\hat{\beta}_0 = \bar{y}$, w związku z tym pierwsze reszty mają postać $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{1}\bar{y}$. W każdym kroku algorytm identyfikuje zmienną x_j najbardziej skorelowaną z aktualną resztą, a następnie dokonuje aktualizacji:

$$\hat{\beta}_j \leftarrow \hat{\beta}_j + \delta_j \quad \text{oraz} \quad \boldsymbol{r} \leftarrow \boldsymbol{r} - \delta_j \cdot \boldsymbol{x_j},$$

gdzie $\delta_j = \varepsilon \cdot sign\left(\text{cov}\left(\mathbf{x}_j, \mathbf{r}\right)\right)$.

Parametr skracania ε jest małą liczbą dodatnią ustalaną arbitralnie. W związku z tym, algorytm wykonuje bardzo dużą liczbę kroków i w praktyce wiele kroków z rzędu może aktualizować ten sam parametr. Oznacza to też, że w początkowych

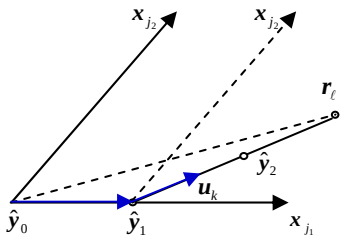
iteracjach wiele współczynników pozostaje równych zero, co jest równoznaczne z selekcją zmiennych. Naturalnym kryterium stopu jest brak korelacji między jakimkolwiek predyktorem a aktualnymi resztami.

3. Metoda LARS

Inspiracją metody LARS, opracowanej przez Efrona i in. [2004], był wcześniej przedstawiony algorytm modelowania addytywnego ze skracaniem. Zasadnicze różnice to aktualizacje na podstawie kilku najbardziej skorelowanych predyktorów oraz to że w każdej iteracji do modelu wprowadzana jest kolejna zmienna, co w praktyce oznacza, że wystarczy przeprowadzić *p* iteracji.

Nazwa ***L**east **A**ngle **R**egre**S**sion* wywodzi się z interpretacji geometrycznej tej metody. Rozważmy zmienne jako wektory w *N*-wymiarowej przestrzeni obiektów. Przez $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}$ oznaczmy predyktory najbardziej skorelowane z wektorem reszt $\boldsymbol{r}$. Wobec tego, że ustawienia początkowe są takie same jak w poprzednim algorytmie, są one standaryzowane. Dalej rozważana będzie podprzestrzeń liniowa $\ell(x_{j_1}, x_{j_2})$ rozpięta przez te wektory (rys.1). Oznaczmy przez $\boldsymbol{r}_\ell$ rzut ortogonalny wektora $\boldsymbol{r}$ na tę podprzestrzeń. Warto nadmienić, że $\boldsymbol{r}_\ell$ jako kombinacja liniowa wektorów $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}$ stanowi rozwiązanie MNK. Przez $\hat{\boldsymbol{y}}_k$ oznaczane są wartości teoretyczne w kolejnych iteracjach (jako punkty w wielowymiarowych przestrzeniach), przy czym $\hat{\boldsymbol{y}}_0 = \boldsymbol{0}$.

W pierwszym kroku wybierany jest predyktor $\boldsymbol{x}_{j_1}$ najbardziej skorelowany z $\boldsymbol{r}$. Aktualizowana będzie wartość parametru β_{j_1}, co geometrycznie oznacza przesunięcie punktu $\hat{\boldsymbol{y}}_0$ do nowej pozycji $\hat{\boldsymbol{y}}_1$ w kierunku wektora $\boldsymbol{x}_{j_1}$. Długość tego przesunięcia jest tak dobrana, by wektor $\boldsymbol{r}_\ell - \hat{\boldsymbol{y}}_1$ leżał na dwusiecznej kąta między wektorami $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}$. Inaczej mówiąc, wektor $\boldsymbol{r}_\ell - \hat{\boldsymbol{y}}_1$ ma wówczas jednakowe korelacje z $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}$. Wektor ten wyznacza kierunek kolejnego przesunięcia punktu wartości teoretycznych (w kolejnej iteracji), a przez $\boldsymbol{u}_k$ w formule (2) oznaczony jest współliniowy z nim wektor jednostkowy. Aktualizacja wartości teoretycznych oznacza zarazem zmianę aktualnego wektora reszt. Długość kolejnego przesunięcia punktu wartości teoretycznych ponownie dobrana jest tak, by wyznaczyć kierunek w przestrzeni trójwymiarowej, tworzący jednakowe kąty z wektorami $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}$ oraz trzecim wektorem reprezentującym kolejną zmienną wprowadzaną do modelu na zasadzie maksymalizacji skorelowania z aktualnymi resztami.



Rys. 1. Geometria metody LARS

Źródło: opracowanie własne na podstawie [Efron i in. 2004].

Formuła aktualizacyjna wygląda następująco:

$$\hat{\boldsymbol{y}}_k = \hat{\boldsymbol{y}}_{k-1} + \gamma \cdot \boldsymbol{u}_k \text{, gdzie: } \boldsymbol{u}_k = A_k \cdot X_k (X_k^T X_k)^{-1} \boldsymbol{1}_k . \tag{2}$$

Symbole $X_k, \boldsymbol{1}_k, A_k$ oznaczają kolejno: macierz zbudowaną z *k* zmiennych objaśniających, *k*-wymiarowy wektor jedynek oraz liczbę rzeczywistą wyrażającą kowariancję między wektorem $\boldsymbol{u}_k$ a zmiennymi wprowadzonymi do modelu (są one równe, gdyż wektor $\boldsymbol{u}_k$ tworzy jednakowe kąty z *k* zmiennymi w modelu). Z kolei wartość γ, decydująca o długości przesunięcia w kierunku wektora $\boldsymbol{u}_k$, ustalana jest jako mniejsza spośród dodatnich wartości:

$$\left\{ \frac{\hat{C} - \hat{c}_j}{A_k - a_j}, \frac{\hat{C} + \hat{c}_j}{A_k + a_j} \right\}, \tag{3}$$

gdzie $\hat{c}_j$ jest korelacją nowo wprowadzanej zmiennej z aktualnymi resztami, $\hat{C}$ maksymalną (co do wartości bezwzględnej) korelacją predyktora z resztami ***r***, natomiast a_j iloczynem skalarnym nowo wprowadzonej zmiennej z wektorem $\boldsymbol{u}_k$.

Po wprowadzeniu wszystkich zmiennych do modelu (a więc w *p*-tym kroku) ostatnie przesunięcie punktu reprezentującego wartości teoretyczne dokonuje się tak, by osiągnąć koniec wektora aktualnych reszt (rozwiązanie MNK).

W wyniku przeprowadzenia tego algorytmu otrzymuje się *p* alternatywnych modeli, z różną liczbą zmiennych objaśniających. Należy więc wybrać optymalny według wybranego kryterium, na przykład błąd predykcji estymowany sprawdzaniem krzyżowym lub C_p Mallowsa (wyniki tej statystyki dostępne są bezpośrednio po zastosowaniu procedury `lars` w programie R). Jest to zarazem równoznaczne z selekcją zmiennych.

4. Badania empiryczne

Przeprowadzony eksperyment miał na celu zbadanie odporności metody LARS na występowanie zmiennych nieistotnych. Zdolność eliminowania zmiennych nieistotnych porównano też z innymi, powszechnie stosowanymi metodami selekcji zmiennych. Badania przeprowadzono dla pięciu zależności funkcyjnych, z których ostatnie dwie pochodzą z pakietu `mlbench` programu R:

Funkcja 1: $y = 300 + x_1 + 0{,}5x_2 + 5x_3 - 2x_4 + e$.

Funkcja 2: $y = x_1 - 3x_3 + x_1x_2 - 0{,}3x_1x_3 + e$.

Funkcja 3: $y = 300 + x_4^3 - x_1^2x_3 + 0{,}5x_4^2 - x_2x_3 - 20x_3x_4 + x_2 + e$.

Funkcja 4: $y = \sqrt{x_1^2 + \left(x_2x_3 - \frac{1}{x_2x_4}\right)^2} + e$.

Funkcja 5: $y = arc\,tg\,\frac{x_2x_3 - \frac{1}{x_2x_4}}{x_1} + e$.

Wartości zmiennych objaśniających losowane były z rozkładów jednostajnych i nie zachodzi między nimi korelacja liniowa. Za każdym razem generowano zbiór uczący oraz testowy. Zbiory testowe miały zawsze liczebność 500, natomiast uczące 500 lub 50. Szum gaussowski $e - N(0, s)$ dodawano tylko w zbiorach uczących, a parametr *s* był ustalany jako 0,1 odchylenia standardowego zmiennej objaśnianej *y*. Do zbiorów dołączano też 10 zmiennych nieistotnych Z_j, których wartości losowane były z rozkładów jednostajnych z ustalonym *min* = 0 oraz *max* losowanym z przedziału od 10 do 1000. Między zmiennymi nieistotnymi Z_j wprowadzano zależności liniowe:

$$Z_2 = 2Z_1 + 0{,}5Z_3 + 5Z_4 + e_2,$$

$$Z_{10} = 10Z_5 + Z_6 + 2Z_7 + 8Z_9 + e_{10},$$

gdzie $e_j - N(0; s_j)$ oraz $s_j = 0{,}3 \cdot sd(Z_j)$ dla $j \in \{2;10\}$.

Do porównania zdolności identyfikacji zmiennych nieistotnych wybrano następujące metody selekcji zmiennych. W klasycznej regresji liniowej (estymacja metodą najmniejszych kwadratów MNK) eliminowano zmienne za pomocą testu istotności współczynników. Taką eliminację powtarzano aż do momentu, gdy wszystkie pozostałe zmienne uznawane były za istotne. Regresję krokową stosowano w obu kierunkach przeszukiwania (dobór zmiennej lub eliminacja), a funkcją kryterium było kryterium AIC. W metodzie LARS oraz LASSO wyboru postaci modelu dokonywano na podstawie błędu predykcji estymowanego za pomocą

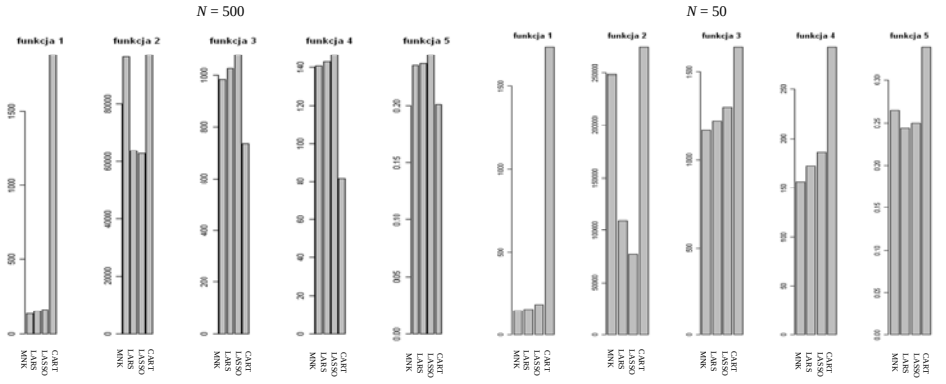
sprawdzania krzyżowego. Dodatkowo stosowano regułę jednego błędu standardowego. W metodzie LASSO sprawdzano 100 wartości progowych *t*. Zastosowano też drzewa regresyjne CART, metodę uznawaną za bardzo odporną na zmienne nieistotne. Drzewa przycinano na podstawie kosztu i złożoności, z uwzględnieniem reguły jednego błędu standardowego [zob. Gatnar 2001].

Tabela 1. Liczby wprowadzeń zmiennych nieistotnych w dziesięciu eksperymentach (przypadek $N = 500$). W nawiasach podano, ile zmiennych nieistotnych wprowadza algorytm w jednym eksperymencie (zakres)

Funkcje	MNK (test t)	Regresja krokowa		LASSO	LARS	CART
		dobór	eliminacja			
Funkcja 1	4 (1-3)	4 (1-2)	6 (1-2)	2 (1-3)	1 (1)	0
Funkcja 2	0	8 (1-3)	8 (1-4)	0	0	0
Funkcja 3	4 (1-3)	5 (1-2)	6 (1-3)	1 (1)	1 (1)	0
Funkcja 4	3 (1-2)	7 (1-2)	8 (1-5)	0	0	0
Funkcja 5	3 (1)	5 (1)	7 (1)	0	0	1 (1)

Źródło: obliczenia własne.

W tabeli 1 przedstawiono wyniki uzyskane w przypadku prób uczących o liczebności 500. Dla każdej funkcji zbiory uczące oraz testowe generowano 10 razy. Zdecydowanie najmniej zmiennych nieistotnych wprowadzają drzewa, a następnie metoda LARS. Najsłabsza jest regresja krokowa. Rysunek 2 (lewe okno) pokazuje, jak na tym tle wyglądają błędy standardowe regresji szacowane na zbiorach testowych. Dla najbardziej skomplikowanych zależności (funkcje 3-5) bezkonkurencyjny był CART (małe błędy i nie wprowadza zmiennych nieistotnych). Dla zależności liniowych i interakcji najskuteczniejsze były LARS i LASSO (w zależności liniowej otrzymywano wprawdzie nieco większe błędy, ale nie wprowadzały tak dużo zmiennych nieistotnych).



Rys. 2. Błędy standardowe regresji szacowane na zbiorach testowych

Źródło: opracowanie własne.

Z kolei tabela 2 przedstawia wyniki uzyskane w przypadku prób uczących o liczebności 50. Wśród modeli liniowych najlepsza okazała się tu eliminacja zmiennych nieistotnych testem t w klasycznej regresji opartej na metodzie najmniejszych kwadratów. Metoda LARS zawodzi, zwłaszcza w przypadku zależności liniowej. Natomiast drzewa wprawdzie nie wprowadzają zmiennych nieistotnych, ale prowadzą do dużych błędów predykcji (rys. 2 prawe okno), co było spowodowane tym, że odrzucały też zmienne istotne. LARS i LASSO są natomiast nadal konkurencyjne w zależności liniowej z interakcjami.

Tabela 2. Liczby wprowadzeń zmiennych nieistotnych w dziesięciu eksperymentach (przypadek $N = 50$). W nawiasach podano, ile zmiennych nieistotnych wprowadza algorytm w jednym eksperymencie (zakres)

Funkcje	MNK (test t)	Regresja krokowa		LASSO	LARS	CART
		dobór	eliminacja			
Funkcja 1	2 (1-2)	5 (1-5)	7 (1-5)	7 (1-4)	9 (1-4)	0
Funkcja 2	2 (1-3)	5 (1-2)	6 (1-5)	5 (1-3)	3 (1-2)	0
Funkcja 3	1 (1)	4 (1-3)	7 (1-3)	3 (1)	2 (1-4)	0
Funkcja 4	4 (1-5)	7 (1-5)	9 (1-5)	1 (2)	4 (1-3)	0
Funkcja 5	3 (1-2)	5 (1-2)	5 (1-3)	5 (1-2)	1 (1)	0*

* brak podziałów

Źródło: obliczenia własne.

4. Wnioski

Metody LARS nie zaleca się stosować do selekcji zmiennych w małych próbach, choć może dać relatywnie dobre rezultaty, gdy zależność funkcyjna jest liniowa z interakcjami. Znacznie lepsze wyniki można wówczas otrzymać, eliminując zmienne nieistotne testem t w klasycznej metodzie najmniejszych kwadratów. Z kolei w dużych próbach LARS jest metodą godną polecenia, zwłaszcza jeśli zależność funkcyjna jest liniowa lub liniowa z interakcjami. W przypadku skomplikowanych zależności funkcyjnych konkurencyjne są jednak drzewa regresyjne.

Literatura

Blum A.L., Langley P., *Selection of relevant features and examples in machine learning*, „Artificial Intelligence" 1997, vol. 97 no. 1-2, s. 245-271.

Efron B., Hastie T., Johnstone I., Tibshirani R., *Least angle regression*, „Annals of Statistics" 2004, 32 (2), s. 407-499.

Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, Wyd. Naukowe PWN, Warszawa 2001.

Guyon I., Gunn S., Nikravesh M., Zadeh L., *Feature Extraction: Foundations and Applications*, Springer, New York 2006.

Guyon I., Elisseeff A., *An introduction to variable and feature selection*, „Journal of Machine Learning Research" 2003, 3, s.1157-1182.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*: *Data Mining, Inferance, and Prediction*, 2nd ed., Springer, New York 2009.

Liu H., Yu L., *Toward integrating feature selection algorithms for classification and clustering*, „IEEE Transactions on Knowledge and Data Engineering" 2005, 17, s. 491-502.

Miller A., *Subset Selection in Regression*, 2nd ed., Chapman and Hall/CRC, Boca Raton, 2002.

Osborne M., Presnell B., Turlach B., *A new approach to variable selection in least squares problems*, „IMA Journal of Numerical Analysis" 2000, 20, s. 389-404.

Tibshirani R., *Regression shrinkage and selection via the lasso*, „J. Royal. Statist. Soc. B." 1996, 58, s. 267-288.

THE ANALYSIS OF LARS METHOD IN FEATURE SELECTION PROBLEM IN REGRESSION

Summary: Feature selection is a typical task of data mining when a researcher looks for an interesting and unsuspected relations in the large data-sets without prior knowledge about the examined phenomenon. In linear regression, the iterative estimation methods can be applied for this purpose (i.e. LARS proposed by Efron et al. [2004]) instead of popular stepwise regression or classical testing of the significance of coefficients. The goal of this paper is to test the abilities of LARS in the identification of irrelevant variables, especially when some of them are collinear. The comparison between some feature selection methods is also given.