Jan E. Zejda

ZJAWISKO WSPÓŁLINIOWOŚCI W ANALIZIE WIELU ZMIENNYCH: PRZYCZYNY, ROZPOZNANIE I MOŻLIWOŚCI ELIMINACJI PROBLEMU*

COLLINEARITY IN MULTIVARIABLE ANALYSIS: CAUSES, DETECTION AND CONTROL MEASURES

Katedra Epidemiologii Śląski Uniwersytet Medyczny w Katowicach

STRESZCZENIE

Artykuł przedstawia definicję, przyczyny i możliwości rozpoznania oraz korygowania zjawiska współliniowości, zniekształcającego wyniki analizy wielu zmiennych (analizy wielowymiarowej). Poza omówieniem danych literaturowych opisujących podstawowe metody postępowania w odniesieniu do wymienionych kwestii w artykule przytoczony jest własny przykład, odwołujący się do wyników analizy zależności pomiędzy występowaniem bólu ramion u pracowników biurowych regularnie stosujących komputery podczas pracy a ich wiekiem, stażem pracy oraz przeciętnym dziennym czasem pracy na stanowisku komputerowym. Przykład wykorzystuje wyniki analizy regresji liniowej i demonstruje obecność zjawiska współliniowości (korelacja pomiędzy dwiema zmiennymi niezależnymi: wiekiem i stażem pracy) oraz jego zniekształcający wpływ na oszacowanie współczynników regresji. Wyniki analizy modelu kompletnego (wiek i staż pracy w modelu) są konfrontowane z wynikami analizy modelu zredukowanego (albo wiek, albo staż pracy w modelu). Ponadto, w odniesieniu do omawianego przykładu, w artykule zaproponowane są praktyczne sposoby identyfikacji zjawiska współliniowości, w tym znaczenie analizy korelacji liniowej oraz tzw. analizy tolerancji.

Słowa kluczowe*: współliniowość, regresja liniowa, rozpoznanie i kontrola współliniowości*

ABSTRACT

The paper reviews principal effects of collinearity on the results of multivariate regression analysis. The discussion focuses on the definition of the problem and on practical means of its recognition, as well as on preventive measures aiming at control of collinearity. In addition to the literature-based review the paper includes the presentation of a case study involving assessment of the relationship between pain in arms and age, years of work, daily duration of work in men and women regularly using personal computers at work. Case-study data were used to show the effect of collinearity (interdependence of two independent variables: age and years of work) on the coefficients of regression in a saturated model, followed by demonstration of the changes resulting from restriction measures (either age or years of work in the model). In addition, in relation to the case-study, the paper shows the results of practical means of detection of collinearity, including analysis of linear correlation and tolerance diagnostics.

Key words: *collinearity, linear regression analysis, detection and control of collinearity*

Rozpowszechnienie statystycznych programów komputerowych umożliwiających szybkie przeprowadzenie analizy wielu zmiennych (analizy wielowymiarowej) jest niezwykłym ułatwieniem pracy epidemiologa. Analizy tego typu umożliwiają ilościową i jakościową ocenę efektu objętego badaniem (zależność Y od X) po uwzględnieniu wpływu innych czynników, co ma istotne znaczenie w przypadku badań nad złożonymi uwarunkowaniami zjawisk zdrowotnych. Ten walor analizy wielu zmiennych jest powszechnie wykorzystywany w analizach ukierunkowanych zarówno na stworzenie tzw. modelu objaśniającego, jak i w analizach ukierunkowanych na stworzenie tzw. modelu predyktywnego (1).

* - W artykule wykorzystano dane zgromadzone w ramach projektu pt. „Ocena stanu zdrowia i podstawowych wskaźników kosztu fizjologicznego pracy u osób zatrudnionych na stanowiskach związanych z obsługą komputera” (Program Wieloletni „Poprawa Bezpieczeństwa i Warunków Pracy”)

Powszechnie dostępne pakiety statystycznej analizy danych oferują liczne wersje procedur reprezentujących modele analizy regresji liniowej i regresji logistycznej. Korzystanie z literatury, także tej, która towarzyszy pakietom statystycznym nie wyklucza możliwości potknięć metodologicznych w trakcie prowadzenia analiz i interpretacji ich wyników. Wśród wielu możliwych przyczyn takich potknięć dość powszechną wydaje się „droga na skróty". Zrozumiałe dążenie do weryfikacji hipotez przy użyciu metod oferujących statystyczne opracowanie efektów zależnych od wielu czynników sprawia, że analizy wielu zmiennych są prowadzone chętnie i dość wcześnie w fazie opracowania wyników badań. Niekiedy są one prowadzone z pominięciem kanonu systematycznej, stopniowej analizy danych, obejmującego w logicznej kolejności - po opisie i diagnostyce rozkładów zmiennych - analizy proste, analizy stratyfikacyjne i analizy wielu zmiennych. Te ostatnie służą weryfikacji wcześniej uzyskanych wyników i nie zastępują prostszych metod. Poza weryfikacją hipotez (analizy konfirmatywne) analizy wielu zmiennych znajdują bodajże jeszcze bardziej powszechne zastosowanie w próbie identyfikacji „godnych uwagi" efektów w zgromadzonym zbiorze danych (analizy eksploratywne). W porównaniu z pierwszym zastosowaniem (analizy konfirmatywne) drugi obszar dociekań (analizy eksploratywne) generuje wyraźnie większe ryzyko popełnienia błędu metodologicznego.

Wśród możliwych pomyłek i błędów znajdują się takie, które wynikają z konstrukcji modelu (równania) analizy wielu zmiennych. Procedury powszechnie dostępne w programach komputerowych umożliwiają testowanie dużej liczby różnorodnych zmiennych niezależnych w jednym modelu. Dodatkowo możliwość skorzystania z funkcji automatycznej selekcji zmiennych niezależnych odznaczających się statystyczną znamiennością może powodować zmniejszenie troski o wprowadzenie do modelu tylko tych zmiennych, wobec których istnieją silne przesłanki natury biologicznej lub statystycznej. Postępowanie to, jak wspomniano bardziej popularne przy prowadzeniu analiz eksploratywnych, może prowadzić do zniekształcenia wyników analizy wielu zmiennych wskutek tzw. zjawiska współliniowości (w języku angielskim: collinearity) (2-4).

Współliniowość jest zniekształceniem wyniku analizy wielu zmiennych wynikającym z wzajemnego powiązania dwóch ilościowych zmiennych niezależnych (X_1 i X_2) włączonych do jednego modelu regresji ($Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$). Powiązanie to może mieć charakter naturalny (wynikający z natury zmiennych, a raczej z natury zjawisk przez te zmienne opisywanych) lub przypadkowy (występujący przypadkowo w analizowanym zbiorze danych). Przykładem zmiennych odznaczających się naturalnym powiązaniem jest masa i wysokość ciała lub stężenie mocznika i kreatyniny w surowicy krwi – w obu przypadkach dwie zmienne są silnie skorelowane. Przykładem innego powiązania (można tu zidentyfikować wpływ procesów społecznych) jest związek pomiędzy wykształceniem matki (liczba lat nauki) i wykształceniem ojca (liczba lat nauki). Wreszcie takie powiązanie może mieć czysto przypadkowy wymiar – na przykład w analizowanym zbiorze danych ciśnienie skurczowe krwi może okazać się dodatnio skorelowane z przeciętnym czasem trwania snu. Niekorzystnym skutkiem współliniowości jest uzyskanie zniekształconych wartości współczynników regresji (b_1, b_2) i wartości 'p' charakteryzujących statystyczną znamienność tych współczynników (duży błąd standardowy współczynników regresji).

Najskuteczniejszym sposobem postępowania wobec zjawiska współliniowości jest profilaktyka. Przede wszystkim należy unikać włączania do jednego modelu tych zmiennych niezależnych, o których wiadomo, że w sposób naturalny są skorelowane. Na przykład, analizując zależność czasu hospitalizacji od stanu czynnościowego nerek należy zrezygnować z włączenia do modelu regresji zmiennych naturalnie skorelowanych, zostawiając w modelu tylko jedną z nich (np. albo „mocznik", albo „kreatynina"). Wpływ każdego ze wskaźników na czas hospitalizacji można oszacować porównując wyniki „modelu z mocznikiem" z wynikami „modelu z kreatyniną". Innym sposobem jest manipulacja zmiennymi. Na przykład analizując zależność występowania astmy dziecięcej od tzw. czynników środowiska rodzinnego można zamiast zmiennej „wykształcenie matki" i zmiennej „wykształcenie ojca" skonstruować jedną, kombinowaną zmienną – wykształcenie rodziców (wyrażone jako średnia liczba lat nauki). Można wreszcie próbować transformacji jednej ze zmiennych, na przykład zmienną „wykształcenie matki" można pozostawić w oryginalnej skali, a zmienną „wykształcenie ojca" przekształcić w dwupoziomową zmienną jakościową. Niezależnie od przedstawionych możliwości logiczną alternatywą jest radykalne rozwiązanie, polegające na pozostawieniu w testowanym modelu tylko jednej z dwóch skorelowanych zmiennych niezależnych.

Wielkość zniekształcenia związanego ze współliniowością dwóch niezależnych zmiennych ilościowych zależy od wartości współczynnika korelacji pomiędzy tymi zmiennymi. W związku z tym zalecaną, prostą metodą identyfikacji zagrożenia współliniowością (także przypadkową) jest wczesne przeprowadzenie analizy korelacji obejmującej wszystkie ilościowe zmienne, typowane jako zmienne niezależne w modelu regresji (2-5). Analiza korelacji zmiennych metodą „każda z każdą" umożliwi ponadto identyfikację zagrożenia wielowspółliniowością (zależność wzajemna trzech lub więcej zmiennych; w języku angielskim: multicollinearity). Spotyka się akceptowaną dość powszechnie

praktykę eliminacji z modelu takiej zmiennej, która pozostaje w korelacji z inną zmienną na poziomie co najmniej r = 0,5 , ale postępowanie to nie wynika z opublikowanych zaleceń. W przypadku niezależnych zmiennych jakościowych o ich wzajemnym powiązaniu informują wyniki testów niezależności (np. testu chi-kwadrat).

Poniższy przykład ilustruje obecność współliniowości i jej wpływ na wyniki analizy wielu zmiennych, a także efekt manipulacji zastosowanych w celu wykluczenia współzmienności. Przykład opracowany został z wykorzystaniem własnej bazy danych i procedur dostępnych w oprogramowaniu statystycznym SAS – wykorzystano funkcje 'proc univariate' i 'proc ttest', w analizie korelacji liniowej funkcję 'proc corr' w modyfikacji Spearman'a, a w analizie regresji liniowej funkcję 'proc reg' (6). W grupie 175 osób zatrudnionych na stanowiskach związanych ze stałym stosowaniem komputera (138 kobiet i 37 mężczyzn) określono nasilenie dolegliwości bólowych ze strony ramion, przy pomocy tzw. skali wizualnej (zakres od 0 do 10 jednostek - U). Zgodnie z protokołem badawczym wśród potencjalnych czynników wpływających na obecność i nasilenie bólu ramion uwzględniono wiek badanych (w latach), ich staż pracy na stanowisku komputerowym (w latach) oraz przeciętny dzienny czas pracy na tym stanowisku (w godzinach). Wyniki pomiaru wymienionych okoliczności przedstawia tabela I.

Tabela I. Nasilenie bólu w obrębie ramion, wiek, staż pracy i przeciętny dzienny czas pracy z komputerem w grupie 175 osób zatrudnionych na stanowiskach komputerowych

Table I. Intensity of pain in the arms, mean age, years of work and average daily time of computer use in 175 persons working at computer stations

Zmienna	Wartość Średnia	Odchylenie standardowe	Zakres
Ból ramion (U)	2,4	2,4	0 – 10
Wiek (lata)	37,3	11,9	19 – 65
Staż pracy (lata)	10,9	6,8	1 – 36
Dzienna praca (min.)	416	82	120 – 720

Wśród okoliczności mogących także wpływać na nasilenie dolegliwości bólowych lub modyfikować wpływ czasu pracy i wieku na te dolegliwości uwzględniono dodatkowo płeć badanych i stosowanie podpórki kończyn górnych podczas pracy z komputerem (tak=113, nie=62). Wyniki analiz prostych wykazały obecność statystycznie znamiennej korelacji liniowej pomiędzy nasileniem bólu ramion i wiekiem (r = 0,17; p = 0,01) oraz stażem pracy (r = 0,17; p = 0,02), przy braku statystycznie znamiennego wpływu dziennego czasu pracy (r = 0,04; p = 0,5). Ponadto, w odniesieniu do nasilenia bólu ramion nie stwierdzono (test t-Studenta) statystycznie znamiennego wpływu stosowania podpórki kończyn górnych podczas pracy (p = 0,3) lub płci (p = 0,4). Wszystkie wymienione potencjalne determinanty bólu ramion zostały włączone do kompletnego modelu regresji, obejmującego następujące zmienne: Ból (U) = płeć (0/1) + podpórka (0/1) + wiek (lata) + staż pracy (lata) + dzienny czas pracy (min.). Wyniki analizy przedstawia tabela II.

Tabela II. Zależność nasilenia bólu ramion od płci, stosowania podpórki przedramion, wieku, stażu i dziennego czasu pracy u 175 osób zatrudnionych na stanowiskach komputerowych – wyniki analizy regresji wielu zmiennych w modelu kompletnym.

Table II. Association of intensity of pain in the arms with gender, use of forearm suport, age, years of work and duration of daily computer use in 175 persons working at computer stations – results of a saturated model of multivariate regression analysis

Parametr	Płeć (0/1)	Podpórka (0/1)	Wiek (lata)	Staż pracy (lata)	Dzienny czas pracy (minuty)
b*	0,393 (0,445)	0,409 (0,383)	0,040 (0,024)	0,025 (0,040)	0,004 (0,002)
p**	0,3	0,2	0,09	0,5	0,04

Objaśnienia: * - współczynniki regresji z ich błędami standardowymi (w nawiasie); ** - statystyczna znamienność współczynników regresji

Legend: * - coefficients of regression and their standard errors (in the brackets); ** - statistical significance of coefficients of regression

Wyniki analizy wielu zmiennych wykazały, że spośród potencjalnych zmiennych objaśniających jedynie wpływ dziennego czasu pracy odznaczał się statystyczną znamiennością (p = 0,04). Biorąc pod uwagę konstrukcję kompletnego modelu można podejrzewać, że wyniki obarczone są zniekształceniem wskutek współliniowości. Podejrzenie to wynika z faktu, że wiek i staż pracy są z reguły silnie skorelowane. To podejrzenie zostało zweryfikowane poprzez wykonanie analizy korelacji. Jej wyniki potwierdziły związek pomiędzy wiekiem i stażem pracy (r = 0,75; p < 0,0001) i słabiej wyrażony związek pomiędzy wiekiem i dziennym czasem pracy (r = -0,31; p<0,0001) oraz stażem pracy i dziennym czasem pracy (r = -0,25; p = 0,0007). Ze względu na silną korelację pomiędzy wiekiem i stażem pracy postanowiono wykluczyć jedną z tych dwóch zmiennych. Z kolei nie przekraczające poziomu 0,5 wartości współczynników r pomiędzy wiekiem i dziennym czasem pracy oraz stażem pracy i dziennym czasem pracy uzasadniały pozostawienie zmiennej „dzienny czas pracy" w modelu. Analizę wielu zmiennych ponowiono dla modeli zredukowanych: (Model I bez stażu pracy): Ból (U) = płeć (0/1) + podpórka (0/1) + wiek (lata) + dzienny czas pracy (min.); (Model II bez wieku): Ból (U) = płeć (0/1) +

podpórka (0/1) + staż pracy (lata) + dzienny czas pracy (min.). Wyniki analiz przedstawia tabela III.

Tabela III. Zależność nasilenia bólu ramion od płci, stosowania podpórki przedramion, wieku albo stażu i dziennego czasu pracy u 175 osób zatrudnionych na stanowiskach komputerowych – wyniki analizy regresji wielu zmiennych w modelu zredukowanym

Table III. Association of intensity of pain in the arms with gender, use of forearm support, age, years of work and duration of daily computer use in 175 persons working at computer stations – results of a reduced model of multivariate regression analysis

Model Zredukowany I: brak zmiennej 'staż pracy' w modelu					
Parametr	Płeć (0/1)	Podpórka (0/1)	Wiek (lata)	Staż pracy (lata)	Dzienny czas pracy (minuty)
b*	0,378 (0,448)	0,431 (0,386)	0,045 (0,016)	Zmienna Usunięta	0,003 (0,002)
p**	0,4	0,2	0,005	Zmienna Usunięta	0,1
Model Zredukowany II: brak zmiennej 'wiek' w modelu					
Parametr	Płeć (0/1)	Podpórka (0/1)	Wiek (lata)	Staż pracy (lata)	Dzienny czas pracy (minuty)
b*	0,331 (0,446)	0,402 (0,386)	Zmienna usunięta	0,074 (0,027)	0,003 (0,002)
p**	0,4	0,2	Zmienna usunięta	0,007	0,09

Objaśnienia: * - współczynniki regresji z ich błędami standardowymi (w nawiasie); ** - statystyczna znamienność współczynników regresji

Legend: * - coefficients of regression and their standard errors (in the brackets); ** - statistical significance of coefficients of regression

Wyniki analiz z wykorzystaniem alternatywnych, zredukowanych modeli wykazały, że zarówno wiek (Model I), jak i staż pracy (Model II) mają statystycznie znamienne znaczenie dla nasilenia dolegliwości bólowych ramion. Ponadto okazało się, że w przypadku zredukowanych modeli wyniki analiz nie potwierdzały już widocznego w modelu kompletnym wpływu dziennego czasu pracy na nasilenie bólu ramion. Porównując wartość współczynników regresji (b) dla wieku (b = 0,045) i stażu (b = 0,074) można nawet przypuszczać, że wpływ wieku jest w tym przypadku mniejszy niż wpływ stażu, ale nie to jest sednem dociekań w omawianym przykładzie. Poza tym interpretacja tego typu musi być ostrożna ze względu na wyraźne zróżnicowanie współczynników zmienności obu zmiennych (wiek, staż). Uzasadnione podejrzenie obecności współliniowości, potwierdzone wynikami analiz korelacyjnych, usprawiedliwiało przeprowadzenie prostej manipulacji polegającej na eliminacji współliniowości poprzez pozostawienie w modelu regresji tylko jednej ze skorelowanych zmiennych niezależnych. Ta procedura okazała się skuteczna

Tabela IV. Diagnostyka współliniowości przy użyciu analizy tolerancji w odniesieniu do kompletnego i zredukowanego modelu regresji, przedstawionego w tekście (w tabeli zawarte są wartości wskaźników tolerancji dla poszczególnych zmiennych).

Table IV. Tolerance analysis in diagnostics of collinearity in saturated and reduced models described in the text (the table shows tolerance measures of independent variables)

Model analizy regresji	Płeć (0/1)	Podpórka (0/1)	Wiek (lata)	Staż pracy (lata)	Dzienny czas pracy (minuty)
Model kompletny	0,98	0,99	0,40	0,43	0,87
Model zredukowany	0,99	0,99	Zmienna Usunięta	0,93	0,93

– w analizach wykorzystujących zredukowane modele potwierdzono zależność bólu ramion i od wieku i od stażu pracy na stanowiskach komputerowych.

Przytoczony przykład ilustruje konsekwencje współliniowości i praktyczny sposób wyeliminowania związanego ze współliniowością zniekształcenia wyników analizy wielu zmiennych. Na gruncie statystycznej interpretacji wyników za podejrzeniem omawianego zniekształcenia przemawia obecność dużych błędów standardowych współczynników regresji dotyczących zmiennych objętych współliniowością, a także duża zmiana współczynników regresji, gdy do modelu wprowadza się lub gdy z modelu eliminuje się jedną ze zmiennych biorących udział w kształtowaniu współliniowości. Warto przy tym odnotować, że zjawisko współliniowości w większym stopniu obciąża oszacowanie współczynników regresji niż ogólną zdolność modelu regresji do wyjaśnienia analizowanych zależności, określaną na podstawie wartości współczynnika dopasowania modelu – statystyki R^2 informującej w jakim stopniu zestaw zmiennych niezależnych w danym modelu wyjaśnia zmienność w zakresie zmiennej zależnej w tym modelu.

Współczesne programy statystyczne oferują automatyczną detekcję skutków współliniowości. Tego typu metody są szczególnie przydatne, gdy występuje zagrożenie ze strony wielowspółliniowości, zwłaszcza przy analizie eksploratywnej. Automatyczne techniki detekcji współliniowości funkcjonują w oparciu o oszacowanie jak dalece zmienność w zakresie jednej zmiennej niezależnej może być wiązana ze zmiennością w zakresie innej zmiennej niezależnej (co jest sednem współliniowości). Oszacowanie to jest prowadzone dla każdej zmiennej w modelu i jest ono możliwe w odniesieniu do zmiennych ilościowych i jakościowych. Wśród stosowanych metod dość powszechnie sięga się do techniki znanej jako „analiza tolerancji" (3,6). Jej wynikiem jest statystyka znana jako „wskaźnik tolerancji", obliczony dla każdej zmiennej. Przydatność „wskaźnika tolerancji" wynika z dość prostej inter-

pretacji jego wartości: po pierwsze - zakres wartości tej statystyki kształtuje się w zakresie od 0 do 1; po drugie - gdy wartość tej statystyki jest większa od 0,5 wówczas odznaczająca się tą wartością zmienna wnosi małe zagrożenie (wielo)współliniowością (zmienna „może być tolerowana" w modelu). Trzeba jednakże uprzedzić, że podana jako decyzyjna wartość „0,5" ma charakter umowny i nie znajduje pokrycia w ogólnie obowiązujących rekomendacjach, a omawiana i analogiczne metody detekcji problemu bywają kontestowane na gruncie rozważań metodycznych (4,7).

Praktyczne znaczenie analizy tolerancji ilustruje wynik tej procedury, zastosowanej w odniesieniu do wcześniej przedstawionego kompletnego modelu regresji wielu zmiennych: Ból (U) = płeć (0/1) + podpórka (0/1) + wiek (lata) + staż pracy (lata) + dzienny czas pracy (min.). Wyniki analizy tolerancji przedstawia tabela IV. Wśród zmiennych niezależnych dwie charakteryzują się „wskaźnikiem tolerancji" nie przekraczającym wartości 0,5 – są nimi „wiek" (0,40) i „staż pracy" (0,43). Można zatem przyjąć, że każda z tych zmiennych jest objęta (wielo)współliniowością ze strony pozostałych zmiennych niezależnych. Najprostszym rozwiązaniem w tym przypadku jest korekta konstrukcji modelu polegająca na usunięciu tej zmiennej niezależnej, która odznacza się najmniejszą wartością 'wskaźnika tolerancji' – tu zmiennej „wiek". Jej rezultatem jest zredukowany model: Ból (U) = płeć (0/1) + podpórka (0/1) + staż pracy (lata) + dzienny czas pracy (min.). W modelu zredukowanym, w przypadku każdej zmiennej niezależnej, wskaźnik tolerancji sięga maksymalnej wartości, co upoważnia do stwierdzenia, że wyniki przeprowadzonej ingerencji są satysfakcjonujące.

Przedstawione powyżej znaczenie, konsekwencja i podstawowe (nie wszystkie) metody kontroli zjawiska współliniowości w analizie wielu zmiennych nie mają wyłącznie wymiaru teoretycznego. Łatwość (techniczna) prowadzenia analiz wielu zmiennych i eksploracja rozbudowanych zbiorów danych sprawiają, że problem ma duży wymiar praktyczny. Co więcej, potencjalne skutki współliniowości rzadko stanowią przedmiot rozważań towarzyszących decyzji o podjęciu analizy wielu zmiennych i rzadko uzupełniają rutynowy kanon postępowania, obejmujący przede wszystkim weryfikację założeń odnośnie normalności, liniowości, niezależności, stabilności wariancji.

Pożądaną metodą kontroli zjawiska współliniowości jest przede wszystkim rozwaga przy konstrukcji modelu regresji. Epidemiologia reprezentuje nauki medyczne, a zatem kierunki dociekań muszą podążać za szeroko rozumianym biologicznym prawdopodobieństwem analizowanych zjawisk. Ten kanon dotyczy także modelowania zależności przyczynowo-skutkowej, które jest wielce uproszczoną próbą zapisu złożonych zjawisk biologicznych. Wiedza na temat mechanizmów biologicznych badanych zjawisk ułatwia dostrzeżenie zagrożenia współliniowością, ale nie wyklucza ujawnienia się tego problemu w trakcie analizy danych. W związku z tym przedstawione powyżej metody identyfikacji i kontroli niekorzystnych skutków współliniowości powinny stanowić integralny element analizy wielu zmiennych.

PIŚMIENNICTWO

1. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989; 79: 340-349
2. Kleinbaum D G, Kupper L L, Muller K E, Applied Regression Analysis and Other Multivariate Methods. Boston: PWS-KENT Publishing Company, 1988: 206-217
3. Chan Y H: Biostatistics 201: Linear regression analysis. Singapore Med J 2004; 45: 55-61
4. Tu Y K, Clerehugh V, Gilthrope M S. Collinearity in linear regression is a serious problem in oral health research. Eur J Oral Sci 2004; 112: 389-397
5. Szklo M, Nieto F J. Epidemiology: Beyond the Basics. Geithersburg: Aspen Publishers, Inc., 2000: 187-190
6. SAS Institute Inc. 2004. SAS OnlineDoc® 9.1.3. Cary, NC, USA
7. O'Brien R.M. A caution regarding rules of thumb for Variance Inflation Factors. Quality and Quantity 2007; 41: 673-690

Otrzymano: 23.03.2009 r.
Zakwalifikowano do druku: 14.05.2009 r.

Adres do korespondencji:
Prof. dr hab. med. Jan E. Zejda
Katedra Epidemiologii - Śląski Uniwersytet Medyczny
Ul. Medyków 18
40-752 Katowice
Tel.: 032 252 3734
E-mail: jzejda@sum.edu.pl