ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 162, 2002

*Dariusz Parys**

WIELOKROTNA TESTOWA PROCEDURA KROCZĄCA W ANALIZIE REGRESJI

STRESZCZENIE. Procedura wielokrotna krocząca jest oparta na tradycyjnych testach z analizy regresji takich jak testy F oraz testy korelacji cząstkowej.

Nowa procedura utrzymuje wielokrotny poziom istotności na poziomie wcześniej ustalonym. Procedura ta nie wymaga nowych rozkładów i tablic wartości krytycznych, a jednocześnie jako procedura wielokrotna w przypadku odrzucenia hipotezy zerowej $H_0 : Cov(X_i, Y) = 0$ pozwala na wykrycie zależności, które i w jakim stopniu ze zmiennych zależnych X są skorelowane ze zmienną zależną Y. Bazą analizy jest liniowy model regresji $Y = X\beta + \varepsilon$.

Słowa kluczowe: liniowy model regresji, testowanie wielokrotne, procedury kroczące

I. WPROWADZENIE

Analiza regresji oparta na modelu regresji ma bardzo szerokie zastosowanie. Wykorzystujemy ją w celu opisu danych, utrzymania kontroli nad zmiennymi zależnymi, a także wyjaśnienia i ustalenia struktury zależności między zmiennymi w modelu.

Literatura dotyczy w większości przypadków zastosowania modelu regresji do przewidywania wyników, predykcji. W artykule skoncentrowano się na wybraniu spośród dużej liczby zmiennych niezależnych tych odpowiednich dla modelu regresji. Wielokrotna analiza regresji jest często używaną metodą statystyczną. Ponieważ wnioskowanie w wielokrotnej analizie regresji jest oparte na modelu, wydaje się bardzo istotne aby model ten był odpowiednio skonstruowany. Jeżeli ograniczymy się do liniowych modeli, problem redukuje się do stwierdzenia, które zmienne powinny być wprowadzone do modelu. Tutaj również

* Dr, Katedra Metod Statystycznych, Instytut Ekonometrii i Statystyki, Uniwersytet Łódzki.

ograniczymy się do liniowych modeli regresji. Jeżeli liczba możliwych „regresorów" wzrasta, pojawia się potrzeba stworzenia technik, które bardziej lub mniej automatycznie wybierają najlepszy model.

Zróbmy kilka podstawowych założeń dla naszych modeli. Po pierwsze składniki losowe są niezależne i mają identyczny normalny rozkład z wartością oczekiwaną zero i skończoną wariancją. Mówimy teraz, że testowany model będzie spełniał warunki tzw. „słabe" warunki w analizie regresji.

II. STRUKTURA ZALEŻNOŚCI

Jak wspomniano we wprowadzeniu, jednym z zastosowań modelu regresyjnego jest ustalenie i wyznaczenie wpływu na zmienną zależną wszystkich możliwych zmiennych niezależnych – każdej po kolei lub grupami. Można oczywiście wykorzystać znane standardowe techniki. Jednak jeżeli dochodzimy do momentu, w którym należy podjąć decyzję co do możliwych związków między zmienna zależną a pewnymi zmiennymi niezależnymi, musimy zwrócić naszą specjalną uwagę na poziom istotności oraz możliwą obecność współliniowości między zmiennymi niezależnymi.

Rozważmy zmienną niezależną Y i L zmiennych niezależnych $X_1, ..., X_L$, które „może" mają jakiś wpływ na Y. Celem analizy jest wykrycie które X -y mają ten wpływ. Ponieważ istnieje możliwość współliniowości między zmiennymi niezależnymi, naszym celem jest utworzenie grupy zmiennych mających wspólny wpływ na Y. Końcowym produktem analizy jest klasyfikacja X-ów na M grup. Wszystkie grupy oprócz M-tej pokazują istotny wpływ na Y. M-ta grupa składa się ze zmiennych nie wskazujących na istotny wpływ na Y. Grupa ta może być pusta. Naszą podstawową hipotezą staje się stwierdzenie, że wszystkie X-y należą do ostatniej grupy. Rozważmy przypadek, gdzie $L = 10$.
Hipoteza zerowa ma postać

$$H_0 : (X_1, X_2, ..., X_{10}). \tag{1}$$

Oznacza to, że żadne z 10 X-ów nie są skorelowane z Y, tzn. $M = 1$.

Po analizie struktura okazuje się być następująca:

$$|X_1|X_4\ X_6|X_5|(X_2\ X_3\ X_7\ X_8\ X_9\ X_{10}) \tag{2}$$

co oznacza, że $M = 4$, zmienne $X_2, X_3, X_7, X_8, X_9, X_{10}$ wykazują brak wpływu na Y, zmienne X_1 i X_5 są indywidualnie skorelowane z Y oraz zmienne X_4

i X_6 są skorelowane z Y, lecz jako zmienne skorelowane ze sobą nie są od siebie oddzielone.

Można oczywiście wyobrazić sobie ukrytą zmienną $X_{4,6}$ która *de facto* jest prawdziwym regresorem, lecz ponieważ $X_{4,6}$ nie jest faktycznie zaobserwowaną zmienną lub też możliwą do zmierzenia, musi być reprezentowana przez jedną lub więcej innych zmiennych.

III. WIELOKROTNA TESTOWA PROCEDURA KROCZĄCA

Głównym problemem jest wskazanie na te możliwe zmienne niezależne, które powinny być włączone do końcowego modelu. W poniższej procedurze jest to wykonywane krok po kroku. Rozpoczynamy od zmiennej, która ma największy wpływ na zmienną zależną, potem rozważamy te, które mają coraz mniejszy wpływ. Ponieważ chodzi nam o własny, „bezpośredni" wpływ danej zmiennej, to badanie związku w drugim kroku i dalej jest zdefiniowane warunkowo względem zmiennej już rozważanej w poprzednich krokach. Kroczący charakter procedury i złożoność problemów wymaga specjalnego rozważania.

W pierwszym kroku rozważamy ogólną zerową hipotezę

$$H_0 : Cov(X_i, Y) = 0 \quad i = 1, ..., L \tag{8}$$

Rezultatem pierwszego kroku powinien być X_i taki, który ma największy wpływ na Y. Hipoteza alternatywna ma postać:

$$H_A : \max |Cov(X_i, Y)| > 0 \qquad i = 1, 2, ..., L \tag{9}$$

Oznacza to, że nawet jeżeli dwie lub więcej zmiennych X są istotnie skorelowane z Y, tylko ta jedna, z największą korelacją, powinna być odszukana (wybrana).

Głównym powodem jest uniknięcie, w tym samym momencie, zmiennych niezależnych z dużą korelacją między sobą. Zgodnie z rozważaną wcześniej strukturą takie zmienne powinny zostać zaklasyfikowane do tej samej grupy i reprezentowane przez jedną z nich. Jeżeli zerowa hipoteza w pierwszym kroku nie jest odrzucona, jest oczywistym fakt, że nie ma istotnego wpływu na Y ze strony żadnej ze zmiennych niezależnych.

Z drugiej strony, jeżeli hipoteza zerowa w pierwszym kroku jest odrzucona, należy postawić pytanie, czy istnieje więcej zmiennych istotnie skorelowanych z Y i jakie są zmienne.

Hipoteza zerowa w kroku drugim ma postać

$$H_0 : Cov(X_i, Y - \hat{\alpha} - \beta_1 X^*) = 0 \tag{10}$$

gdzie X^* jest zmienną niezależną wybraną w kroku pierwszym oraz $\hat{\alpha}$ i $\hat{\beta}_1$ są otrzymane metodą najmniejszych kwadratów. To znaczy, korelacja jest mierzona między resztą z modelu w kroku pierwszym i pozostałymi zmiennymi niezależnymi odpowiednio.

Hipoteza alternatywna ma postać:

$$H_A : \max |Cov(X_i, Y - \hat{\alpha} - \hat{\beta}_1 X^*)| > 0 \tag{11}$$

i tak dalej, wybierając tylko jedną zmienną X istotnie skorelowaną z Y.

W drugim i następnych krokach pojawia się pytanie jak formować grupy, o których wspomniano wcześniej. Jeżeli zmienna niezależna, powiedzmy X_2, jest istotnie skorelowana z Y w kroku pierwszym, lecz nie jest skorelowana najsilniej i korelacja ta znika lub co najmniej staje się nieistotna w kroku drugim, wtedy X_2 wybrana zostaje jako ta, która nie ma wpływu na Y.

Maksymalna liczba kroków (lub zerowych hipotez) równa się liczbie możliwych regresorów L. Procedura zatrzymuje się i problem jest rozwiązany, gdy to po raz pierwszy nie jest odrzucona.

Można zauważyć, że hipotezy zerowe różnią się od siebie z kroku na krok. Trudno jest z tego powodu znaleźć formalne przedstawienie tego problemu w „języku" hipotez.

Niech

$$H_0^k : Cov(X_i, Y - \hat{\alpha} - \hat{\beta}_1 X_1^* - \beta_2 X_2^* - \ldots - \hat{\beta}_{k-1} X_{k-1}^*) = 0 \tag{12}$$

gdzie:

k – numer kroku, $i \in I_{k-1}$,

I_{k-1} – zbiór indeksów dla tych zmiennych X, które nie były najbardziej istotne w żadnym z kroków 1, 2, ..., k-1

X_j^* – zmienna najbardziej istotna w kroku j-tym.

Hipoteza alternatywna ma postać:

$$H_0^k : \max Cov|(X_i, Y - \hat{\alpha} - \hat{\beta}_1 X_1^* - \beta_2 X_2^* - \ldots - \hat{\beta}_{k-1} X_{k-1}^*)| > 0 \tag{13}$$

W procedurze tej możliwe decyzje podejmowane po każdym kroku są następujące:

1) Jeżeli H_0 nie jest odrzucona, procedura zatrzymuje się;

2) Jeżeli H_0 jest odrzucona, to zmienna z największą korelacją X_k^* jest wprowadzona do modelu;

3) Jeżeli zmienna $X_i \neq X_{k-1}^*$ jest istotna w kroku $(k-1)$-ym, natomiast nie jest istotna w kroku k-tym, to jest umieszczona w grupie razem z X_{k-1}^* i oznaczamy ją przez X_k^1 itd.

Jeżeli już po pierwszym kroku H_0 nie jest odrzucona, to jest oczywiste, że żadna ze zmiennych X nie ma wpływu na Y. Jeżeli procedura zawierała co najmniej kilka kroków, wtedy rezultaty można przedstawić następująco:

$$| X_1^* X_1^1 ... | X_2^* X_2^1 ... | ... | \tag{14}$$

Model zbudowany jest z tych zmiennych, które okazały się najbardziej istotne.

Podstawą analizy jest model

$$Y = X\beta + \varepsilon \tag{15}$$

gdzie

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1L} \\ 1 & x_{21} & x_{22} & \cdots & x_{2L} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nL} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Główną ideą wielokrotnej testowej procedury kroczącej jest wyselekcjonowanie tych zmiennych niezależnych, które są najbardziej istotnie skorelowane z Y i stworzenie nowego modelu:

$$Y = X^* \beta^* + \varepsilon \tag{16}$$

gdzie X^* – macierz obserwacji na tych zmiennych, które dały istotny wpływ na zmienną Y.

W pierwszym kroku nie ma wprowadzonych żadnych zmiennych i stąd

$$X_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Aby rozpocząć testowanie, X_1 rozszerzamy o jedną kolumnę:

$$X_{1,i} = \begin{bmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{in} \end{bmatrix} \quad i = 1,\ldots,L.$$

Estymatorem $\beta_{1,i}$ w modelu $Y = X_{1,i}\beta_{1,i} + \varepsilon$ jest $b_{1,i} = (X_{1,i})(X'_{1,i})^{-1}(X'_{1,i}Y)$.

Sumę kwadratów $SSR_{1,i} = b'_{1,i} \cdot X_{1,i} \cdot Y$ porównujemy z SSR_1 – początkową sumą kwadratów.

Obliczenia są dokonywane L razy, raz dla każdego możliwego regresora, dając L różnic w SSR, które są porównane do sumy kwadratów reszt. $SSE = SST - SSR = Y'Y - bX'Y$.

Jeżeli H_0 jest prawdziwe, wtedy

$$F_{1,i} = \frac{SSR_0 - SSR_{1,i}}{MSE} \sim F^{1,n-L-1}, \tag{17}$$

gdzie MSE – estymator σ^2 reszt, $n - L - 1$ jest liczbą stopni swobody.

Zmienna X może należeć do następujących kategorii:

I) zmienne bez wpływu na Y,

II) zmienne należące do grupy wspólnego wpływu na Y,

III) zmienne mające swój własny wpływ na Y.

Największy błąd popełniamy uznając zmienną z kategorii jako istotną, stąd zmienna z największą wartością F, czyli dająca największy przyrost SSR powinna być włączona do modelu.

Wartością krytyczną dla testu wielokrotnego utrzymującego poziom istotności poniżej ustalonego poziomu α jest $F_{\alpha/L}^{1,n-L-1}$.

Oczywisty jest fakt, że $P(F_{\alpha/L}^{1,n-L-1} \mid H_0) \le \frac{\alpha}{L}$ prawdopodobieństwo otrzymania jednego błędnego stwierdzenia o istotności $F_{1,i}$ jest nie większe niż α.

Istotnie:

$$P\left(\bigcup_{i=1}^{L} F_{1,i} > F_{\alpha/L}^{1,n-L-1} \mid H_0\right) \le \sum_{i=1}^{L} P\left(F_{1,i} > F_{\alpha/L}^{n-L-1} \mid H_0\right) = L \cdot \frac{\alpha}{L} = \alpha \tag{18}$$

Testowanie jest oparte na $\max F_{1,i}$, $i = 1,\ldots,L$. Nie zmienia to istoty sprawy, gdyż

$$P\left(\max F_{1,i} > F_{\alpha/L} \mid H_0\right) = 1 - P\left(\bigcap_{i=1}^{L} F_{1,i} \le \ldots \mid H_0\right) = P\left(\bigcup_{i=1}^{L} F_{1,i} > \ldots \mid H_0\right) = \alpha \tag{19}$$

Reguła decyzyjna po kroku pierwszym polega na zatrzymaniu procedury, jeżeli $\max F_{1,i} \le F_{\alpha/L}^{1,n-L-1}$ i na kontynuowaniu, jeżeli zachodzi przeciwna nierówność.

Ogólna procedura testowa ma postać:

1) po kroku k-tym procedura zatrzymuje się jeżeli $F_{k,i} < F_{\alpha/L-k+1}^{1,n-L-1}$ dla wszystkich i, gdzie i jest indeksem tych zmiennych, które nie zostały włączone do modelu w kroku 1, 2, ..., (k–1)-tym;

2) Jeżeli $\max F_{k,i} \ge F_{\alpha/L-k+1}^{1,n-L-1}$, $k = 1,2,\ldots$ wtedy odpowiedni X jest włączony do modelu;

3) Jeżeli $\max F_{k-1,i} \ge F_{\alpha/L-k+1}^{1,n-L-1}$, $F_{k,i} < F_{\alpha/L-k+1}^{1,n-L-1}$, $k = 1,2,\ldots$ oraz $(n-2)\left(\frac{r^2}{1-r^2}\right)^2 \ge F_{\alpha}^{1,n-2}$, gdzie r jest współczynnikiem korelacji cząstkowej między i-tą zmienną i zmienna włączoną w (k–1)-tym kroku, wtedy i-ta zmienna jest w tej samej grupie co X_{k-1}^{*}.

IV. WNIOSKI

Metoda zasugerowana w tym artykule pozwala na wykrycie tych zmiennych niezależnych X_i, które są najbardziej istotnie skorelowane ze zmienną zależną Y i stworzenie z nich nowego modelu regresji $Y = X^{*}\beta^{*} + \varepsilon$. W procedurze tej nie

ma potrzeby wprowadzenia testów wymagających od nas wiedzy o rozkładach. Dzięki połączeniu metod tradycyjnych, jak test F oraz test korelacji cząstkowych oraz nieklasycznej procedury wielokrotnej kroczącej uzyskano efekt w postaci utrzymywania wielokrotnego poziomu istotności na poziomie (lub w przybliżeniu) wcześniej ustalonego α, czego nie gwarantowały klasyczne procedury, dające często niejednoznaczne rozwiązania.

BIBLIOGRAFIA

Holm S. (1977), *Sequentially Rejestive Multiple Test Procedures*, „Statistical Research Report", **1**.

Miller R. G. Ir (1980), *Simultaneous Statistical Inference*, 2nd ed: Springer Verlag, New York.

Thompson M. L. (1978), *Selection of Variables in Multiple Regression: Part I. A Review and Evaluation. Part II. Chosen procedures, computations and examples*, Inst. Stat. Rev., **46**, 1–19, 129–146.

Dariusz Parys

THE MULTIPLE STEPWISE PROCEDURE IN REGRESSION ANALYSIS

The multiple procedure for stepwise regression analysis presented in this paper is based on traditional methods, such as F-test and test of partial correlations. This procedure, having multiple testing character, keeps the multiple significance level at a predetermined value, at least approximately.

This approach a way of dealing with, and reporting includes the dependencies among the explanatory variables, including their impact on the dependent one the procedure suggested in this paper does not introduce any new tests of call for any new distributions.