RUCH PRAWNICZY, EKONOMICZNY I **SOCJOLOGICZNY**
Rok XLIV — zeszyt 2 — 1982

PETR HEBÁK, JIRI KŘOVÁK

WYBÓR NAJLEPSZEGO PODZBIORU ZMIENNYCH NIEZALEŻNYCH DO OPISOWEGO MODELU REGRESYJNEGO

I. WSTĘP

Zwolennik stosowania analizy regresyjnej w najróżniejszych dziedzinach, głównie zaś na polu ekonomii, natrafi na samym wstępie rozwiązania konkretnego zadania na dwa problemy, których przezwyciężenie jest warunkiem koniecznym dla powodzenia analizy. Jednym z tych problemów jest odpowiedni wybór zbioru zmiennych niezależnych. Celem mniejszej pracy [1] jest w tym względzie z jednej strony podanie pewnych informacji, a z drugiej podanie klucza wyboru spośród wchodzących w grę metod.

Badanie relacji między wielkościami ekonomicznymi jest stale przedmiotem zainteresowania tych, którzy zajmują się możliwościami modelowania zjawisk i procesów ekonomicznych. Analiza regresji, której się do tych celów używa najczęściej, rozwija się nieustannie od roku 1885, kiedy to Francis Galton zajmował się po raz pierwszy prostymi regresji w czasopiśmie Instytutu Antropologicznego. Schemat klasyczny, kiedy zmiany jednej losowej wielkości są poza przypadkowymi błędami wyjaśniane zmianami tzw. zmiennych niezależnych, a których wartości są kontrolowanie przez eksperymentatora, działa tylko w pewnych warunkach niemierzalnym i trudnym do uchwycenia przypadkowymi składnikiem, który reprezentuje jakiś zbiorczy wpływ nie branych pod uwagę czynników wpływających na omawianą zmienną zależną oraz przy pewnych dalszych założeniach co do kształtu funkcji regresji, co do ilości parametrów tej funkcji w stosunku do wielkości próby, a także co do

[1] W opracowaniu wykorzystano następujące prace: L. Cyhelský i in., *Analiza zjawisk ekonomicznych za pomocą regresji korelacji,* Praca badawcza nr 67, VÜSEI, Praga 1975; N. R. Draper, H. Smith, *Applied Regression Analysis,* New York 1966; M. A. Efroymson, *Multiple Regression Analysis,* in: A. Ralston, H. S. Wilf, *Mathematical Methods for Digital Computers,* New York 1960; A. S. Goldberger, *Econometric Theory,* New York 1964; J. Likes, P. Hebák, *Analiza regresyjna zjawisk ekonomicznych,* Praca badawcza nr 49, VÜSEI, Praga 1973.

charakteru zmiennych zależnych. Niektóre z tych „klasycznych" założeń bywają w sferze ekonomicznej nieuzasadnione. Zmienne niezależne są przeważnie wielkościami losowymi nie są to więc zmienne, które eksperymentator wybiera. Między tymi objaśniającymi wielkościami istnieją najróżniejsze stosunki, które nie bywają „czysto" funkcyjnymi, by można było niektóre zmienne zastąpić pozostałymi, lecz są one wielkościami korelacyjnymi i powodują mniejszą dokładność szacunków regresyjnych oraz komplikują działanie poszczególnych zmiennych niezależnych. Także założenie stałej dyspersji zmiennych niezależnych dla różnych kombinacji wartości zmiennych niezależnych jest w tej sytuacji często nieuzasadnione. Jeśli chodzi o dane uzyskane z obserwacji wielkości w czasie, to przeważnie założenie niezależności poszczególnych obserwacji okaże się niewystarczające. W sytuacji, kiedy zmienne niezależne są wielkościami losowymi, będącymi w innych zadaniach zmiennymi zależnymi, założenie klasyczne jest niewystarczające m. in. dlatego, że zmienne niezależne i element losowy nie są wielkościami przypadkowymi i niezależnymi, a szacunki parametrów nie są asymptomatycznymi bezstronnymi, a więc i stałymi. Przedstawione trudności, wynikające z niedotrzymania „klasycznych" warunków, wymagają innego podejścia do rozwiązywania problemu regresji, nowych metod respektujących wyżej wymienione realia i według możliwości wykorzystujących wszystkie informacje, które są do dyspozycji w danym zadaniu. Mimo wagi powyższych problemów naszym dzisiejszym celem nie jest wskazanie możliwości rozwiązania, ani nie chcemy też przedstawić dróg, którymi kroczy badanie tych zagadnień. Przedmiotem niniejszej pracy jest wprawdzie inny problem, ale jego rozwiązanie jest pośrednio związane ze słusznością klasycznych założeń. Idzie nam o część problemu poszukiwania „najlepszej" funkcji regresji, której istotą jest określenie najlepszego podzbioru zmiennych niezależnych. Chodzi więc o kwestię decyzji, które ze wszystkich wchodzących w grę zmiennych niezależnych wpływają decydująco na zmiany zmiennej zależnej przy danym, przeważnie, linearnym kształcie funkcji regresji i przy dotrzymaniu pewnych założeń. Słowo wstępne o trudnościach, wynikających z bezzasadności niektórych założeń przy stosowaniu analizy regresji w sferze ekonomii, miało czytelnika tylko uprzedzić, że podane dalej metody można stosować tylko wtedy, jeśli uda się jednocześnie rozwiązać niektóre dalsze kwestie lub w sytuacjach, kiedy metody te będą tylko częścią kompleksowego rozwiązania regresji, zawierającego oprócz rozwiązania klasycznego również metody alternatywne.

W pracy J. Likesa i P. Hebáka [2] dokonaliśmy szczegółowej analizy niektórych ogólnych, rzeczowych, logicznych i statystycznych zasad, którymi kierujemy się w pracy z materiałem liczbowym, i których musimy

[1] J. Likes, P. Hebák, *Analiza regresyjna zjawisk ekonomicznych.*

przestrzegać przy poszukiwaniu „najodpowiedniejszej" funkcji regresji. Z braku miejsca nie będziemy ich wszystkich tu przedstawiać, ale ustalmy przynajmniej te, które uważamy za bezsporne.

l) Zbiór zmiennych niezależnych, ich forma przyporządkowania w funkcji regresji, jak i sam kształt funkcji regresji muszą być całkowicie zgodne z teorią ekonomiczną, z doświadczeniami praktycznymi oraz z rzeczowo-logiczną istotą badanego problemu. Bez właściwego sformułowania problemu, bez rzeczowej znajomości problematyki i bez logicznej analizy nie można odpowiednio podejść do rozwiązania żadnego zadania z zastosowaniem regresji.

2) Prostsza funkcja, zależnie od ilości parametrów i kształtu funkcji regresji, jest zarówno z punktu widzenia interpretacji wyników jak i z punktu widzenia przestrzennej, czasowej stabilności funkcji, zawsze lepsza od funkcji bardziej skomplikowanej.

3) Funkcja regresji, która lepiej odpowiada danej reprezentacji (lub danym reprezentacjom), tj. np. funkcja o mniejszej rezydualnej sumie kwadratów odchyleń wartości rzeczywistych i wyliczonych, jest zawsze lepsza, o ile nie dojdzie do sprzeczności z zasadą podaną na pierwszym i drugim miejscu.

4) Dokładność szacunków dokonanych na podstawie funkcji regresji, uwzględniająca informacje dodatkowe, błędy pomiaru, heteroskedastycydę, autokorelację, multikolinearność, kształt rozkładów warunkowych oraz niektóre dalsze fakty, będzie zawsze większa niż w przypadku klasycznego schematu Gaussa.

Nasze podejście do szukania „najlepszego" podzbioru zmiennych niezależnych jest podejściem wyłącznie empirycznym, a do stosującego podane metody należy respektowanie wspomnianych już, a także dalszych zasad stosowania analizy regresji oraz niepodchodzenie do rozwiązania zadania w sposób mechaniczny. Mechaniczna aplikacja analizy regresji kompromituje cały system metodyczny i prowadzi do braku zaufania do metod matematyczno-statystycznych.

II. WNIOSKI OGÓLNE O PARAMETRACH MODELU LINIOWEGO REGRESYJNEGO

Rozważane dalej metody, poszukujące najlepszej grupy zmiennych niezależnych, zakładają, że zmiany jednej wyjaśnianej zmiennej zależnej określają zmiany k objaśniających zmiennych (niezależnych) $k=1, 2, ..., n$. Zmienna zależna ma charakter wielkości losowej, a wartości zmiennych zależnych są wybierane lub zależne są również od przypadku, ale wtedy nie są zależne od wpływów nie branych pod uwagę. Wpływy (czynniki) nie brane pod uwagę ogólnie stanowią w modelu składnik stochastyczny (element zakłóceniowy) i tylko założenia probabilistyczne co do tego elementu oraz niektóre inne pozwalają sprowadzić problem mie-

rzenia zależności do znanego matematyczno-statystycznego problemu szacunku rozkładu prawdopodobieństwa.

Załóżmy funkcję regresji liniowej, która zarówno z punktu widzenia parametrów jak i zmiennych niezależnych przybiera postać:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki},$$

dla $i=1, 2, \ldots, n$, gdzie: η_i — jest hipotetyczną funkcją regresji, przedstawiającą przebieg średnich wartości zmiennej zależnej Y, dla których kombinacje wartości zmiennych niezależnych $x_1, x_2, \ldots, x_k$; $\beta_0, \beta_1, \ldots, \beta_k$ to nieznane parametry funkcji regresji. Modelem wyjściowym będzie zespół równań liniowych w postaci:

$$y_i = \eta_i + \varepsilon_i, \quad \text{dla} \quad i = 1, 2, \ldots, n,$$

gdzie: ε_i to i-ta wartość nieuchwytnego składnika stochastycznego.

Ów zespół równań można zapisać w postaci macierzy:

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

gdzie: **Y** jest wektorem pionowym $(n \cdot 1)$ badanych wartości zmiennej zależnej, **x** jest macierzą $(n \cdot [k+1])$ rozważanych wartości zmiennych niezależnych, **ß** jest wektorem pionowym $([k+] \cdot 1)$ nieznanych parametrów, a s jest wektorem pionowym $(n \cdot 1)$ nieuchwytnych wartości składnika stochastycznego. W warunkach modelu klasycznego (2), (4) należy na podstawie n obserwacji zmiennej Y i zminnych $x_1, x_2, \ldots, x_k$ szacunkowo ustalić nieznane parametry $\beta_0, \beta_1, \ldots, \beta_k$. Można stosunkowo łatwo pokazać ((2), (4)), że we wispomnianych warunkach $\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ jest najlepszym bezstronnym szacunkiem wektora **ß**, a $\hat{\mathbf{Y}} = \mathbf{x}'_* \cdot \mathbf{b}$ najlepszym bezstronnym szacunkiem uwarunkowanej średniej wartości Y i konkretnej wartości y_* dla danej kombinacji wartości $x_{1*}, x_{2*}, \ldots, x_{k*}$

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k-1} = \frac{\text{RSČ}}{n-k-1} = \frac{\text{CSČ} - \text{TSČ}}{n-k-1},$$

gdzie:

$\text{RSČ} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ tzw. rezydualna suma kwadratów,

$\text{TSČ} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ tzw. teoretyczna suma kwadratów,

$\text{CSČ} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ tzw. całkowita suma kwadratów, przy czym CSČ=TSČ+RSČ.

Bezstronnym szacunkiem rozproszenia składnika stochastycznego, $S(b_j) = s\sqrt{c_{jj}}$ dla $j=0, 1, 2, \ldots, k$, gdzie: s to rezydualne odchylenie standardowe, a c_{jj} jest diagonalnym elementem (macierzy $\mathbf{c} = (\mathbf{x}'\mathbf{x})^{-1}$,

jest bezstronnym szacunkiem standardowego błędu szacunku *j*-tego parametru.

Do oceny stopnia wpływu zmiennych niezależnych na zmienną zależną używa się w tej sytuacji dobrze znanych prostych, cząsteczkowych oraz wielokrotnych koeficjentów korelacji, cząsteczkowych testów *t*-Studenta oraz zmodyfikowanych ogólnych i cząsteczkowych testów *F*-Fischera.

O ile między zmiennymi. niezależnymi nie ma zależności, można cząstkowe współczynniki korelacji oraz cząstkowe testy *t* zastosować do oceny indywidualnego wpływu zmiennych niezależnych. W przeciwnym razie może dojść do sprzeczności między wartościami cząstkowymi, współczynnikami korelacji, a wartością ogólnego współczynnika korelacji, a także może dojść do sprzeczności między wnioskami wysnutymi na 'podstawie cząstkowych testów *t,* a wnioskami wysnutymi na podstawie ogólnego testu *F*.

Ogólny współczynnik korelacji jak i ogólny test *F* umożliwa ocenę całkowitego wpływu wszystkich zmiennych niezależnych razem wziętych, podczas gdy cząstkowe testy *F* stosuje się do oceny przyrostu (ubytku) teoretycznej sumy kwadratów przez dołączenie (wykluczenie) określonej zmiennej lub grupy zmiennych. Z wyjątkiem metody „wszystkich możliwych regresji" wszystkie dalej podane metody korzystają z faktu, że przez dołączenie nowej zmiennej (lub jej funkcji) nie może dojść do spadku teoretycznej sumy kwadratów i że wtedy:

$$\Delta \mathrm{TS\check{C}} = \mathrm{TS\check{C}}_k - \mathrm{TS\check{C}}_{k-h} \geqslant 0,$$

gdzie: $\mathrm{TS\check{C}}_k$ (TSK_k) jest teoretyczną sumą kwadratów dla *k* zmiennych w funkcji regresji, a $\mathrm{TS\check{C}}_{k-h}$ (TSK_{k-h}) jest teoretyczną sumą kwadratów dla *k—h* zmiennych w funkcji regresji *h=1, 2, ..., k—1*. Można powiedzieć, że o ile empiryczna funkcja regresji nie przebiega przez wszystkie warunkowe średnie zmiennej zależnej dla różnych kombinacji wartości zmiennych niezależnych, to przez zwiększenie liczby zmiennych w funkcji regresji dochodzi do zwiększenia teoretycznej sumy (kwadratów (do zmniejszenia rezydualnej sumy kwadratów), a więc i do podniesienia wielokrotnego współczynnika korelacji, ewentualnie (jeśli chodzi o funkcję nieliniową z punktu widzenia zmiennych niezależnych) i do zwiększenia indeksu korelacji. Do tego zwiększenia dojdzie niezależnie od tego czy dołączenie nowej zmiennej jest rozsądne i czy można lub nie można go uzasadnić. Fakt ten wymaga pewnej ostrożności przy interpretacji wyników regresji, ponieważ moglibyśmy bardzo łatwo dojść do funkcji regresji, „z zadowalającym" współczynnikiem korelacji, bez względu na rzeczową stronę badanego stosunku. Tak oto pracowicie uzyskana funkcja regresji, „doskonale" odpowiadająca konkretnym danym, może być zupełnie bezużyteczna z punktu widzenia innej próbki, może nie posiadać żadnej stabilności przestrzennej i/lub czasowej. Niebezpie-

czeństwo to uwzględniają tzw. cząstkowe testy F, które oceniają (testują) założenie czy dołączenie (wykluczenie) określonej zmiennej lub grupy zmiennych wpływa poważnie na teoretyczną sumę kwadratów ze statystycznego punktu widzenia. Przy obaleniu tego założenia twierdzimy, z uprzednio wybranym dość dużym prawdopodobieństwem, że przyłączenie (wykluczenie) stosowanej grupy zmiennych znacznie ulepsza (pogarsza) rozważaną funkcję regresji. Całe postępowanie przeważnie się kończy, gdy nie można już żadnej zmiennej dołączyć lub wykluczyć z funkcji regresji zależnie od tego, o którą metodę chodzi: czy o metodę „wstecznej eliminacji zmiennych", czy o metodę „stopniowego dołączenia zmiennych", czy też o jej ulepszający wariant, (uwzględniający kolejność wprowadzanych zmiennych, tj. metodę regresji krokowej {stepwise regression).

Do badań wielkości ekonomicznych w literaturze zaleca się metodę regresji etapowej (stagewise regression), która nie daje wprawdzie najlepszych bezstronnych ocen, ale usuwa negatywny wpływ multilinearności. Ciekawe jest, że poszczególne metody nie prowadzą do zupełnie jednakowych wyników, ale do wyników dość podobnych. Wraz z rozwojem techniki obliczeniowej można się spodziewać, że metoda „wszystkich możliwych regresji" stanie na czele zainteresowań użytkowników, ponieważ niektóre nowsze, dalej przedstawione kryteria zezwalają na stosunkowo łatwy wybór odpowiedniego podzbioru zmiennych. Jak dotąd jest to metoda niewykonalna z punktu widzenia techniki obliczeniowej.

III. METODY STOSOWANE DO POSZUKIWAĆ „NAJLEPSZEGO" PODZBIORU ZMIENNYCH

W tej części pracy podamy opis poszczególnych metod najczęściej stosowanych do poszukiwań „najlepszego" podzbioru zmiennych, opis programów, którymi wobec niektórych z tych metod dysponuje Instytut VUSEI oraz doświadczenia, które zdobyliśmy przy ich aplikacji. Rozpatrzony te metody według wskazanej poniżej kolejności:

1) stopniowe dołączanie zmiennych,
2) metoda regresji krokowej,
3) wsteczna eliminacja zmiennych,
4) metoda regresji etapowej,
5) wszystkie możliwe regresje.

S t o p n i o w e d o ł ą c z a n i e z m i e n n y c h. Obliczanie metodą stopniowego dołączania zmiennych wychodzi z macierzy prostych współczynników korelacji między zmiennymi niezależnymi nawzajem oraz między zmienną zależną i niezależną w postaci:

$$\mathbf{R}=\begin{vmatrix} r_{11} & r_{12} & \cdots & r_{1k} & r_{1y} \\ r_{21} & r_{22} & \cdots & r_{2k} & r_{2y} \\ r_{31} & r_{32} & \cdots & r_{3k} & r_{3y} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} & r_{ky} \\ r_{y1} & r_{y2} & \cdots & r_{yk} & r_{yy} \end{vmatrix}$$

Jako pierwszą zmienną „kandydującą" do włączenia do równania regresji wybiera się tę, która wykazuje najwyższy prosty współczynik korelacji ze zmienną zależną. Zmienną włączamy do regresji, jeśli ma ona statystycznie znaczny wkład w wyjaśnienie rozproszenia zmiennej zależnej. Z każdym dalszym ogólnym krokiem wylicza się cząstkowe współczynniki korelacji między zmienną zależną a wszystkimi (zmiennymi niezależnymi, dotąd nie wyłączonymi do równania regresji, z powodu już wcześniej włączonych zmiennych. Jako „.kandydata" do włączenia do regresji rozważa się zmienną wykazującą najwyższy cząstkowy współczynnik korelacji ze zmienną zależną, o ile odpowiadający jej diagonalny element inwertowanej macierzy **R** nie jest zbyt bliski zera. Każdy krok w inwersji macierzy **R** odpowiada włączeniu lub wyłączeniu zmiennej. Jeśli dokonujemy inwersji według *i*-tego elementu diagonalnego, odpowiada to włączeniu *i*-tej zmiennej do regresji, o ile nie jest ona jeszcze zawarta w równaniu, lub wyłączeniu *i*-tej zmiennej z regresji, jeśli była ona w którymś z poprzednich kroków do regresji włączona. Wtedy jest więc dana zmienna w przybliżeniu liniową (kombinacją zmiennych już w regresji zawartych. Chociaż dalej omówimy wstępnie sposoby obliczania poszczególnych metod, nie będziemy się zajmować szczegółami algorytmu obliczeniowego [3].

Podstawowe zasady algorytmu są podobne dla metod stopniowego dołączania zmiennych, regresji krokowej i wstecznej eliminacji zmiennych. Zmienna brana pod uwagę przy włączaniu do regresji jest włączana w tym przypadku, jeżeli znacznie zwiększa TSČ, a więc wtedy, gdy cząstkowy test *F* jest znaczny. Jeśli ów test jest znaczny, zmienną włączamy do regresji, a za pomocą cząstkowych współczynników korelacji poszukuje się dalszej wchodzącej w grę zmiennej. Jeśli ów test nie jest znaczny, procedura jest zakończona. W Instytucie VÚSEI jest do dyspozycji program FORWARD, którym prowadzi się obliczenia metodą stopniowego dołączania zmiennych. Zużycie czasu maszynowego jest następujące: zadanie nr 1 o rozmiarze 13 obserwacji oraz 4 zmienne niezależne zajmują 12 sekund, zadanie nr 2 o rozmiarze 185 obserwacji i 6 zmiennych niezależnych zajmuje około 17 sekund. Wejścia programu FORWARD są podobne jak wejścia dalszych programów, o których będzie jeszcze mowa. Wprowadza się:

[3] N. R. Draper, H. Smith, *Applied Regression Analysis;* M. A. Efroymson, *Multiple Regression Analysis.*

M — liczba obserwacji,

N — liczba zmiennych (zmiennych niezależnych + zmienna zależna),

F 1 — tablicowa wartość dla cząstkowego testu *F*,

TOL — wartość parametru, który nie pozwoli rozważać włączenia tych zmiennych, które są w przybliżeniu liniową kombinacją już włączonych zmiennych,

X (I, J) — wartość obserwacji poszczególnych zmiennych.

Program ma następujące wejścia:

— średnie wartości poszczególnych zmiennych,

— macierz **R** prostych współczynników korelacji,

— z każdym krokiem wartość cząstkowego testu *F* oraz kolejny numer włączanej (wprowadzanej zmiennej),

— otrzymane (rezultatywne) równanie regresji oraz testy *t* poszczególnych współczynników regresji,

— procent wyjaśnionego rozproszenia,

— całkowity test *F* równania wynikowego.

M e t o d a r e g r e s j i k r o k o w e j. Metoda regresji krokowej jest w zasadzie zbieżna z metodą stopniowego dołączania zmiennych z jedną tylko zmianą, a mianowicie, że z każdym krokiem badamy zasadność nie tylko wprowadzenia (dołączenia) ostatniej zmiennej, ale i zasadność zachowania, utrzymania wszystkich zmiennych niezależnych, wprowadzonych już w poprzednich krokach. Bierzemy więc pod uwagę wpływ, jaki ma wprowadzenie nowej zmiennej na położenie zmiennych wprowadzonych do regresji w krokach poprzednich. Może się zdarzyć, że w wyniku interkorelacji zmiennych niezależnych, po wprowadzeniu określonej zmiennej, okaże się, że zachowanie innej zmiennej, wprowadzonej w którymś z uprzednich kroków, będzie zbyteczne.

W każdym kroku metody obliczeniowej testuje się więc cząstkowym testem *F* z jednej strony zasadność wprowadzenia do regresji następnej zmiennej oraz z drugiej strony zasadność zachowania już wprowadzonych zmiennych. W danym kroku obliczamy statystykę *F* dla każdej zmiennej tak, jakby była ostatnią, która weszła do regresji i porównujemy z wybraną wartością tablicową. O ile cząstkowy test *F* daje wynik mało istotny, zmienna zostaje wykluczona, w przeciwnym wypadku pozostaje w regresji. Zasadność wprowadzenia następnej zmiennej testuje się zgodnie z metodą stopniowego dołączania zmiennych. Procedura zostaje zakończona, kiedy już żadna zmienna nie wpływa poważnie na TSČ i żadna zmienna nie jest w regresji nadmierna.

Do celów metody regresji krokowej mamy w Instytucie VÚSEI program STEPWISE. Zużycie czasu maszynowego jest przy tej metodzie nieznacznie wyższe niż przy metodzie stopniowego dołączania zmiennych. Największe zadanie, które rozwiązywaliśmy programem STEPWISE, miało rozmiar 1168 obserwacji, 9 zmiennych niezależnych i wy-

magało 3 minut 24 sekund czasu maszynowego. Dane wejściowe wprowadzane były bezpośrednio z dysku magnetycznego, na którym jest w CDB przy Federalnym Urzędzie Statystycznym ułożona gałęziowa baza danych ASIS przemysł.

Metoda STEPWISE daje przeważnie identyczne wyniki jak metoda FORWARD, tylko w kilku mało praktycznych aplikacjach doszło do wykluczenia już wcześniej wprowadzonej zmiennej. Wtedy wyniki obu metod się różnią.

Wejścia programu STEPWISE różnią się od wejść programu FORWARD tylko tym, że wprowadzany jest jeszcze parametr *F* 2, czyli tablicowa wartość testu *F* dla wykluczenia już wprowadzonej zmiennej, której wartość jest taka sama lub minimalnie mniejsza niż krytyczna wartość *F 1* dla wprowadzenia zmiennej.

Wyjścia są zbieżne, z tą tylko różnicą, że w każdym kroku drukuje się wartość testowego kryterium dla włączenia i wykluczenia już wprowadzonych zmiennych oraz liczbę porządkową zmiennej, która jest w danym kroku bądź wprowadzona lub wykluczona.

W s t e c z n a e l i m i n a c j a z m i e n n y c h. Metoda obliczeniowa wychodzi z macierzy odwrotnej (inwersyjnej) do macierzy parowych współczynników korelacji, co odpowiada równaniu regresji, które zawiera wszystkie zmienne z kompletnego zbioru rozważanych zmiennych niezależnych. Za pomocą ogólnego testu *F* stwierdzimy czy przynajmniej jedna zmienna niezależna ma znaczny wpływ na zmienną zależną, W przypadku znacznego ogólnego testu *F* rozpoczniemy właściwą eliminację.

„Kandydatem" do wykluczenia z równania regresji jest ta zmienna, która ma najmniejszy udział w wyjaśnieniu rozproszenia zmiennej zależnej. Taką zmienną z regresji wykluczymy, o ile cząstkowy test *F* jest nieznaczny, a więc jeśli ubytek (spadek) TSČ spowodowany jej wykluczeniem jest bez znaczenia. Postępujemy tak dotąd, dokąd wykluczenie którejkolwiek z pozostałych zmiennych obniżyłoby znacznie teoretyczną sumę kwadratów.

W Instytucie VÚSEI dla metody wstecznej eliminacji zmiennych jest do dyspozycji program BACKWARD. Zużycie czasu maszynowego jest tu poniekąd większe niż w dwu poprzednich metodach, co wynika stąd, że należy najpierw dokonać całkowitej inwersji macierzy parowych współczynników korelacji (tj. wprowadzić do regresji wszystkie zmienne z kompletnego zbioru zmiennych niezależnych), a dopiero potem uruchomić algorytm wyboru zmiennych. Zadanie nr 1 o rozmiarach (13 × 4) wymagało 13 sekund czasu maszynowego, zadanie nr 2 o rozmiarach (185 × 6) — 23 sekund.

Wejścia programu BACKWARD są zbliżone do programu STEPWISE z tą różnicą, że nie wprowadza się parametru *F 1* i *TOL*. Na wyjściu,

po wydrukowaniu macierzy korelacji, pojawia się ogólny test *F* kompletnego równania regresji, a w każdym kroku zaś wartość cząstkowego testu *F* dla wykluczenia zmiennej oraz (Liczba porządkowa zmiennej wykluczanej. Poza tym wyjście jest zbieżne z wyjściem programu STEPWISE. Metoda BACKWARD daje przeważnie wyniki zbieżne z wynikami metod FORWARD i STEPWISE, lecz czasem mogą się one też różnić.

M e t o d a r e g r e s j i e t a p o w e j . Metoda ta zalecana jest przez niektórych autorów do rozwiązywania problemów w dziedzinie ekonomii, ponieważ, jak już było powiedziane, podobno eliminuje wpływ wzajemnej zależności, która właśnie w tej dziedzinie jest silna.

Podstawowa myśl jest następująca: jako pierwszą obejmiemy regresją zmienną zależną, która jest najbardziej skorelowana ze zmienną zależną, Wyliczamy szacunkowe wartości zmiennej zależnej z równania regresji zawierającego tylko tę zmienną. Obliczamy wartości rezydualne jako różnicę zaobserwowanych wartości zmiennej zależnej.

W drugim kroku te wartości rezydualne stają się nową zmienną zależną. Następną wprowadzoną zmienną jest ta, która jest najsilniej skorelowana z nową zmienną zależną — z wartością rezydualną z pierwszego kroku. Przeprowadzimy regresję parową wartości rezydualnych na zmienną przeznaczoną do wprowadzenia. Obliczone równania łączymy razem.

W każdym następnym kroku wstępuje do regresji zmienna najbardziej skorelowana z wartością rezydualną *z* kroku poprzedniego. Proces ten kończy się, kiedy żadna z pozostałych zmiennych nie jest już znacznie skorelowana z aktualnymi wartościami rezydualnymi. Wynikiem (metody regresji etapowej są szacunki wypaczone, ale często wydatniejsze niż szacunki uzyskane metodą najmniejszych kwadratów. Równanie końcowe jest inne niż to, które byśmy uzyskali dla tych samych zmiennych przy bezpośrednim zastosowaniu metody najmniejszych kwadratów,

W Instytucie VÚSEI do metody regresji etapowej jest wykorzystywany program STAGEWISE. Zużycie czasu maszynowego jest w przybliżeniu identyczne jak przy metodach FORWARD i STEPWISE. Na wejściu programu STAGEWISE nie wprowadza się parametrów *F 1* i *F 2*, parametr *TOL* reprezentuje tu krytyczną wartość prostego współczynnika korelacji dla odpowiedniej liczby stopni swobody. Na wyjściu po wydrukowaniu macierzy prostych współczynników korelacji pojawi się w każdym kroku wartość współczynnika korelacji między kolejną zmieniną zależną (kolejnymi bieżącymi wartościami rezydualnymi) a „kandydatami" do włączenia (wprowadzenia) oraz liczba porządkowa wprowadzonej (włączonej) zmiennej. Poza tym wejścia na EMC zbliżone do poprzednich programów.

Wszystkie wymienione metody są zaprogramowane w języku FOR-

TRAN, a wszystkie wyliczenia przeprowadzono za ich pomocą na komputerze CDC 3300 w CDB przy Federalnym Urzędzie Statystycznym lub za pośrednictwem stacji „200 User Terminal" umieszczonej w Instytucie VÜSEI i podłączonej do komputera CDC 3300 we VVD Bratysława.

W s z y s t k i e m o ż l i w e r e g r e s j e. Metoda ta zakłada, że wyliczymy wszystkie możliwe funkcje regresji, które można utworzyć z kompletnego zbioru rozważanych zmiennych niezależnych. Jeśli jest ich k, wtedy można zestawić 2^k różnych równań regresji. Np. dla $k=10$, co nie jest szczególnie wysoką liczbą, $2^k=1024$ równań regresji. Zużycie czasu maszynowego, potrzebnego do ich wyliczenia, byłoby ogromne. Istnieją wszak różne sposoby, które skutecznie podchodzą do wszystkich równań w *ten* sposób, by równanie regresji obliczone w danym kroku nie było gorsze (według uprzednio ustalonego kryterium) niż regresja wyliczona w kroku poprzednimi, np,:

$$C_p=\frac{\mathrm{RS\check{C}}_p}{S^2}-n+2p, \quad \text{lub} \quad J_p=\mathrm{RS\check{C}}_p\frac{n+p}{n-p}.$$

Ponieważ w tej chwila nie dysponujemy żadnym programem do metody wszystkich możliwych regresji, ograniczyliśmy się tyko do jej bardzo zwięzłej charakterystyki, chociaż idzie o metodę, która obecnie jest w centrum zainteresowania fachowców danej problematyki.

IV. WNIOSKI KOŃCOWE

W niniejszym artykule próbowaliśmy wysunąć pewne problemy, które związane są z zastosowaniem analizy regresji w sferze ekonomii. Z całego kompleksu problemów wybraliśmy stosunkowo wąski problem wyboru odpowiednich zmiennych niezależnych do wielokrotnych relacji regresji. Przedstawiliśmy kilka metod, które mechanicznie, według określonego kryterium, wybierają „najlepszy" zbiór zmiennych. Na zakończenie należy zwrócić uwagę, że słowo „(najlepszy" ma tu znaczenie warunkowe. Podzbiór zmiennych, który na podstawie wyników danej metody zdecydujemy się wprowadzić do funkcji regresji, jest wprawdzie najlepszy z punktu widzenia danej metody, ale to też jest wszystko. Według naszych doświadczeń jest w zasadzie obojętne, które metody wyboru zastosujemy. Jeśli jednak moglibyśmy dać którejś z nich pierwszeństwo, prawdopodobnie zalecilibyśmy metodę regresji krokowej. Dalej musimy podkreślić, że wybór najlepszego podzbioru zmiennych niezależnych jest raczej problemem rzeczowym niż statystycznym i nie można go rozwiązywać opierając się tylko na wynikach omówionych tu metod. Mogą one dać rozwiązującemu użyteczny kompas, ale ostateczna decyzja należy do statystyka, który się konkretnym zadaniem zajmuje.

Z języka czeskiego tłumaczyła Leokadia Bielówka

CHOICE OF THE BEST SET OF INDEPENDENT VARIABLES

Summary

The authors tries to throw a new light on some problems related to application of the regression analysis in the field of economics. A Choice of independent variables for the multiple regression has been selected from among many problems. Few methods are presented in which according to the specific criteria the "best" set of variables is chosen. The notion "the best" has here a conditional meaning. After many experiments the authors concluded that a choice of the best subset of independent variables is a problem merits rather thar of statistical character. It cannot be only solved on the basis of formal criteria used in the presented methods, it can be of greait value but the last decision taken on the ground of formal and merits reasons belongs to a researcher.

Tłumaczył Tadeusz Kowalski