Krzysztof Jajuga

UWAGI O STOSOWANIU LINIOWEGO MODELU REGRESYJNEGO W BADANIACH EKONOMICZNYCH

1. Uwagi dotyczące teoretycznych koncepcji liniowego modelu regresyjnego

Liniowy model regresyjny jest tym narzędziem statystycznym,które jest najczęściej stosowane w badaniach praktycznych, szczególnie w ekonomii. Nieczęsto jednak badacze stosujący liniowy model regresyjny zastanawiają się nad jego istotą. Skupiają oni swą uwagę albo na aplikacjach, albo na technikach estymacyjnych. Należy jednak pamiętać, że liniowy model regresyjny może być różnie pojmowany i interpretowany w konkretnych sytuacjach badawczych.

Problemowi temu w literaturze polskiej najwięcej uwagi poświęca Zbigniew Czerwiński w artykule [2], a także w monografii [1]. W literaturze zagranicznej jest również nieco wzmianek w kilku monografiach. Dokonamy obecnie podziału funkcjonujących koncepcji liniowego modelu regresyjnego ze względu na kilka kryteriów.

I tak, jednym z kryteriów podziału może być rodzaj zależności odzwierciedlanych przez liniowy model regresyjny.

Między zmiennymi modelu regresyjnego mogą występować dwa rodzaje zależności. Jedną z nich jest tzw. zależność przyczynowo-skutkowa (jakkolwiek ta nazwa budzi u niektórych badaczy mieszane uczucia), tzn. badacz na podstawie wiedzy merytorycznej stwierdza, iż pewna zmienna wpływa na inną zmienną. Drugą jest zależność typu korelacyjnego, tzn. na podstawie wiedzy merytorycznej stwierdza się brak tego wpływu, jakkolwiek współczynnik korelacji między zmiennymi może sugerować coś innego.

Wobec tego liniowy model regresyjny może być konstruowany dla zbadania zależności przyczynowo-skutkowych lub dla zbadania zależności korelacyjnych. Oczywiście w zagadnieniach ekonomicznych dominuje ten pierw-

szy typ badania. Drugi typ badania ma charakter pomocniczy, a wykorzystywany jest np. przy pewnych metodach estymacji.

Innym kryterium podziału jest koncepcja funkcji regresji. Funkcja regresji może być interpretowana na trzy różne sposoby:

- jako funkcja aproksymująca zmienną objaśnianą za pomocą kombinacji liniowej zmiennych regresyjnych,
- jako funkcja zadana przez badacza na postawie wiedzy merytorycznej o badanym zjawisku, tzn. badacz wie, że zmienne regresyjne oddziałują na zmienną objaśnianą w sposób liniowy, nie zna jedynie współczynników kombinacji liniowej,
- jako funkcja będąca warunkową wartością oczekiwaną łącznego rozkładu zmiennej objaśnianej i zmiennych regresyjnych.

W badaniach ekonomicznych dominuje pierwsza koncepcja funkcji regresji, jakkolwiek pożądane byłoby, aby model ekonometryczny odpowiadał koncepcji drugiej.

Wreszcie, trzecim kryterium podziału koncepcji liniowych modeli regresyjnych, które zostanie omówione, jest „struktura stochastyczna" modelu.

Wyróżnić można tutaj trzy podstawowe rodzaje modeli (dla uproszczenia przedstawimy tu modele z jedną zmienną regresyjną):

Model A.

$$y = f(x) + e$$

gdzie y,x,e są zmiennymi nielosowymi.

W modelu tym pary (x_i,y_i) interpretowane są jako wyniki obserwacji, natomiast e jest błędem aproksymacji. Model ten odpowiada koncepcji regresji jako funkcji aproksymującej zmienną objaśnianą za pomocą kombinacji liniowej zmiennych regresyjnych.

Model B1.

$$Y = f(x) + \epsilon$$

gdzie x jest zmienną nielosową, Y,ϵ są zmiennymi losowymi.

Model B2.

$$Y_i = f(x) + \epsilon_i \quad i=1,\ldots,n$$

gdzie x jest zmienną nielosową, Y_i,ϵ_i są zmiennymi losowymi $(i=1,\ldots,n)$.

Interpretacja tych modeli może być dwojakiego rodzaju:

- pary (x_i,y_i) są wynikami eksperymentu polegającego na tym, że na funkcję ustalonej wielkości zmiennej nielosowej x_i działa zmienna losowa ϵ (lub ϵ_i) i otrzymuje się na wyjściu realizację y_i zmiennej losowej Y (lub Y_i);
- pary (x_i,y_i) są obserwacjami (tzn. wynikami eksperymentu pasywnego), realizacjami zmiennej nielosowej x i zmiennej losowej Y (lub zmiennych losowych $Y_1,Y_2,\ldots,Y_n$).

Z pierwszą interpretacją spotykamy się głównie w doświadczalnictwie, a z drugą w badaniach ekonomicznych.

W zagadnieniach ekonomicznych z modelem B1 mamy do czynienia w przypadku, gdy obserwacje dane są w postaci szeregów przekrojowych, tzn.dysponuje się n realizacjami tej samej zmiennej losowej Y. Z modelem B2 mamy do czynienia wtedy, gdy obserwacje dane są w postaci szeregów czasowych, tzn. dysponuje się po 1 realizacji n zmiennych losowych: $Y_1, Y_2, \ldots, Y_n$.

Model C1.

$$Y = f(X) + \epsilon$$

gdzie Y, X, ϵ są zmiennymi losowymi.

Model C2.

$$Y_i = f(X_i) + \epsilon_i \qquad i=1,\ldots,n$$

gdzie Y_i, X_i, ϵ_i są zmiennymi losowymi $(i=1,\ldots,n)$.

Również tutaj interpretacja modeli może być dwojakiego rodzaju: jako wynik eksperymentu lub wynik obserwacji.

Nieco uwagi poświęcimy obecnie interpretacji funkcji regresji jako warunkowej wartości oczekiwanej rozkładu łącznego zmiennej objaśnianej i zmiennych regresyjnych.

Dla modelu z jedną zmienną regresyjną:

$$Y = E(Y|X=x) + \epsilon$$

W liniowym modelu regresyjnym zakłada się liniowość tej warunkowej wartości oczekiwanej, tzn.

$$E(Y|X=x) = \alpha X + \beta$$

W praktyce nie zawsze tak musi być. Spośród rozkładów dwuwymiarowych wektora losowego (X,Y) nie dla wszystkich warunek liniowości zachodzi.Własność ta zachodzi np. dla dwuwymiarowych rozkładów: normalnego, Studenta, Pearsona II typu, Pareto, beta, F oraz dla różnych postaci dwuwymiarowego rozkładu gamma. Nie zachodzi natomiast np. dla dwuwymiarowego rozkładu Cauchy'ego, Rhodesa oraz dla różnych postaci dwuwymiarowego rozkładu wykładniczego.

Nie zawsze liniowość warunkowej wartości oczekiwanej wystarcza, aby zasadne było stosowanie klasycznych procedur. Przykładowo, rozważmy pewną postać dwuwymiarowego rozkładu gamma, o funkcji gęstości danej wzorem:

$$f(x,y) = \frac{a^{p+q}}{\Gamma(p)\Gamma(q)} x^{p-1} (y-x)^{q-1} e^{-ay}$$

gdzie $y > x > 0$, $a,p,q > 0$, przy czym a,p,q są parametrami tego rozkładu. Warunkowa wartość oczekiwana dla tego rozkładu jest równa:

$$E(Y|X=x) = x+q/a$$

Przykład ten poddaje w wątpliwość zasadność stosowania klasycznych procedur (np. metody najmniejszych kwadratów), skoro a priori wiadomo, że współczynnik regresji równy jest 1.

2. Uwagi dotyczące aplikacji liniowego modelu regresyjnego w badaniach ekonomicznych

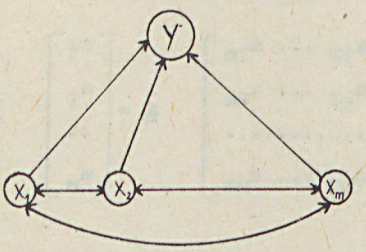
Jedną z przyczyn nieskuteczności stosowania liniowego modelu regresyjnego w zagadnieniach ekonomicznych jest fakt, iż przez wielu badaczy stosowany jest on niemal automatycznie, w abstrakcji od wiedzy merytorycznej o badanym zjawisku. Zdarza się również tak, iż badacz stosujący liniowy model regresyjny zapomina o celu przeprowadzanego badania.

W literaturze formułuje się trzy podstawowe cele stosowania liniowego modelu regresyjnego w badaniach ekonomicznych. I tak, model ten może być konstruowany w celu:

- opisu mechanizmu kształtowania się zjawisk ekonomicznych (tzw. cel poznawczy),
- przewidywania dalszego przebiegu zjawisk ekonomicznych w czasie (tzw. cel predyktywny),
- sterowania dalszym przebiegiem zjawisk ekonomicznych w czasie (tzw.cel decyzyjny).

Z punktu widzenia kryterium odniesienia w czasie cel poznawczy różni się od celu predyktywnego i decyzyjnego. Cel poznawczy dotyczy analizy zależności między zjawiskami w przeszłości, natomiast cel predyktywny i decyzyjny dotyczą wykorzystania modelu w przyszłości. Zatem w tych dwóch ostatnich przypadkach dokonuje się wnioskowania wykraczając poza okres, z którego dane były wykorzystywane przy budowie modelu. Z kolei w zadaniu poznawczym model jest wykorzystywany (do analizy) tylko dla okresu, z którego pochodziły dane stosowane przy budowie modelu.

Obecnie uwagę poświęcimy właśnie zadaniu poznawczemu. W tym celu rozważmy model jednorównaniowy, tzn. taki, w którym żadna ze zmiennych objaśniających nie jest objaśniana przez inną. Model taki można symbolicznie przedstawić w postaci grafu.



W grafie tym Y oznacza zmienną objaśnianą, $X_1, X_2, \ldots, X_m$ - zmienne regresyjne, natomiast jednostronna strzałka oznacza występowanie zależności przyczynowo-skutkowej, a dwustronna strzałka zależności typu korelacyjnego.

Z punktu widzenia celu poznawczego w modelu najważniejsza jest interpretacja ocen parametrów. Pojawiają się tu wady modelu, w sytuacji, gdy jest więcej niż jedna zmienna regresyjna. Mówi się wtedy, że ocena parametru stosjącego przy zmiennej regresyjnej wskazuje, o ile w przybliżeniu wzrośnie wartość zmiennej objaśnianej, gdy ta zmienna regresyj-

na wzrośnie o jednostkę, a pozostałe zmienne regresyjne nie ulegną zmianie. Praktyczna użyteczność takiej interpretacji jest żadna, ponieważ zmiana wartości jednej zmiennej regresyjnej pociąga za sobą z reguły również zmiany wartości innych zmiennych regresyjnych, wywołane istnieniem zależności typu korelacyjnego.

Oznaczmy przez b_i $(i=1,\ldots,m)$ współczynnik wskazujący o ile zmieni się zmienna objaśniana Y gdy zmienna regresyjna X_i wzrośnie o jednostkę. Pomija się tu zatem założenie o braku zmiany wartości pozostałych zmiennych regresyjnych.

Wyznaczamy następujące modele regresyjne opisujące zależności między zmiennymi (bez zmniejszania ogólności rozważań zakładamy, że operujemy obserwacjami scentrowanymi):

$$\hat{Y} = a_1X_1 + a_2X_2 + \ldots + a_mX_m \tag{1}$$

$$\hat{X}_j = a_{ij}X_i \quad i=1,\ldots,m \quad j=1,\ldots,m \quad i\neq j \tag{2}$$

Zauważmy, że zmiana wartości X_i $(i=1,\ldots,m)$ o jednostkę powoduje (na podstawie grafu oraz równań (1) i (2)):

1. zmianę wartości Y o a_i,
2. zmianę wartości X_j $(j=1,\ldots,m,\ i\neq j)$ o a_{ij}.

Z kolei zmiana wartości X_j $(j=1,\ldots,m,\ i\neq j)$ o a_{ij} powoduje zmianę wartości Y o $a_{ij}\,a_j$.

Wobec tego:

$$b_i = a_i + \sum_{\substack{j=1\\ j=i}}^{m} a_{ij}\,a_j = \sum_{j=1}^{m} a_{ij}\,a_j \quad i=1,\ldots,m$$

przy czym $a_{ii} = 1$ $(i=1,\ldots,m)$.

Wprowadźmy oznaczenia macierzowe:

$$\underline{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \quad \underline{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{bmatrix} \quad \underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{bmatrix}$$

Stąd:

$$\underline{b} = \underline{A}\,\underline{a}. \tag{3}$$

Oznaczmy macierz obserwacji zmiennych regresyjnych przez $\underline{X}$, a wektor obserwacji zmiennej objaśnianej przez $\underline{y}$, przy czym

$$\underline{X} = [\underline{x}_1\ \underline{x}_2\ \cdots\ \underline{x}_m]$$

Wobec tego:

$$\underline{A} = \{a_{ij}\}, \quad \text{gdzie } a_{ij} = (\underline{x}_i^T\underline{x}_i)^{-1}\,\underline{x}_i^T\underline{x}_j$$

$$\underline{a} = (\underline{X}^T\,\underline{X})^{-1}\,\underline{X}^T\,\underline{y} \tag{4}$$

Macierz $\underline{A}$ można przedstawić w postaci:

$$\underline{A} = \underline{C}\ \underline{X}^T\ \underline{X} \tag{5}$$

gdzie:

$$\underline{C} = \begin{bmatrix} (\underline{x}_1^T\ \underline{x}_1)^{-1} & 0 & \dots & 0 \\ 0 & (\underline{x}_2^T\ \underline{x}_2)^{-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & (\underline{x}_m^T\ \underline{x}_m)^{-1} \end{bmatrix} \tag{6}$$

Podstawiając (4), (5) i (6) do (3) otrzymujemy:

$$\underline{b} = \underline{C}\ \underline{X}^T\ \underline{X}(\underline{X}^T\ \underline{X})^{-1}\ \underline{X}^T\ \underline{y} = \underline{C}\ \underline{X}^T\ \underline{y} = \begin{bmatrix} (\underline{x}_1^T\ \underline{x}_1)^{-1}\ \underline{x}_1^T\ \underline{y} \\ (\underline{x}_2^T\ \underline{x}_2)^{-1}\ \underline{x}_2^T\ \underline{y} \\ \dots \\ (\underline{x}_m^T\ \underline{x}_m)^{-1}\ \underline{x}_m^T\ \underline{y} \end{bmatrix} \tag{7}$$

Jak widać z (7), wartości współczynników b_i $(i=1,\dots,m)$ są równe współczynnikom regresji modeli z jedną zmienną regresyjną.

Rozważania te wskazują na przewagę modeli z jedną zmienną regresyjną, jeśli badanie jest prowadzone jedynie w celu poznawczym. Model z wieloma zmiennymi regresyjnymi nie przynosi w tym zakresie więcej informacji. Jeśli analiza merytoryczna jest przeprowadzona poprawnie i m zmiennych regresyjnych wyczerpuje zbiór zmiennych merytorycznie istotnych, to należy budować m odrębnych modeli z jedną zmienną objaśniającą.

Literatura

[1] Czerwiński Z.: Matematyczne modelowanie procesów ekonomicznych, PWN, Warszawa 1982.

[2] Czerwiński Z.: O różnych koncepcjach regresji, w: Przegląd Statystyczny, 3-4/1980, s. 221-238.