METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
X, 2009, str. 286 - 292

UOGÓLNIONA MIARA DOPASOWANIA W MODELU LINIOWYM

Wojciech Zieliński
Katedra Ekonometrii i Statystyki, SGGW
Nowoursynowska 159, PL-02-767 Warszawa
wojtek.zielinski@statystyka.info

Streszczenie: W modelu regresji liniowej stosowane są dwie miary wpływu zmiennych niezależnych na zmienną zależną: współczynnik determinacji oraz współczynnik korelacji cząstkowej. W pracy zaproponowane jest uogólnienie tych miar i skonstruowana jest miara wpływu grupy zmiennych niezależnych z wyłączeniem wpływu pozostałych zmiennych niezależnych.

Słowa kluczowe: miara dopasowania, współczynnik determinacji, korelacja cząstkowa, analiza regresji

Załóżmy, że obserwacje Y_i pewnej zmiennej losowej możemy przedstawić w postaci

$$Y_i = \beta_1 x_{1i} + \cdots + \beta_i x_{pi} + \varepsilon_i, \quad i = 1, \dots, n \tag{*}$$

gdzie x_{ki}, $k = 1, \dots, p$, $i = 1, \dots, n$, są znane, $\beta_1, \dots, \beta_p$ są nieznanymi współczynnikami regresji, natomiast ε_i, $i = 1, \dots, n$, są zmiennymi losowymi o rozkładach normalnych o zerowej wartości oczekiwanej i wariancji σ^2.

Jednym z pytań stawianych w analizie modeli (*) jest pytanie o wpływ zmiennych niezależnych na cechę Y. W zastosowaniach wykorzystuje się w zasadzie dwie miary: miarę łącznego wpływu wszystkich zmiennych niezależnych oraz miarę wpływu pojedynczych zmiennych (z eliminacją wpływu pozostałych). Jest to powszechnie znany współczynnik determinacji oraz współczynnik korelacji cząstkowej. Ścisłe określenia tych pojęć można znaleźć w bardzo bogatej literaturze poświęconej analizie regresji (w spisie literatury podanych znaleźć kilka takich książek). Podane miary niestety nie udzielają odpowiedzi na pytania o wpływ grup wybranych zmiennych niezależnych na zmienną zależną, np. jak zmierzyć wpływ zmiennej x_1 oraz x_2 z wyłączeniem pozostałych zmiennych? W dalszym ciągu zaproponowana jest taka miara. Pokazano również, że współczynnik determinacji oraz współczynnik korelacji cząstkowej są szczególnymi przypadkami tej ogólniejszej miary.

W dalszych rozważaniach znacznie wygodniej jest się posługiwać zapisami macierzowymi. Model liniowy (*) w zapisie macierzowym ma postać

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

gdzie $\boldsymbol{Y}' = [Y_1, \dots, Y_n]$ jest wektorem obserwacji, $\boldsymbol{\beta}' = [\beta_1, \dots, \beta_p]$ jest wektorem nieznanych parametrów, $\boldsymbol{\varepsilon}' = [\varepsilon_1, \dots, \varepsilon_n]$ jest wektorem błędów losowych oraz

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{p1} \\ x_{12} & \cdots & x_{p2} \\ \vdots & \cdots & \vdots \\ x_{1n} & \cdots & x_{pn} \end{bmatrix}$$

jest macierzą eksperymentu. Zgodnie z poczynionymi założeniami wektor losowy $\boldsymbol{Y}$ ma n-wymiarowy rozkład normalny:

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$$

Niech teraz $\boldsymbol{X} = [\boldsymbol{X}_1 \vdots \boldsymbol{X}_2]$, gdzie macierze $\boldsymbol{X}_1$ oraz $\boldsymbol{X}_2$ są macierzami o wymiarach odpowiednio $n \times (p-q)$ i $n \times q$. Niech wektor parametrów $\boldsymbol{\beta}$ będzie podzielony odpowiednio na dwa podwektory $\boldsymbol{\beta}_1$ i $\boldsymbol{\beta}_2$ zgodnie z podziałem macierzy $\boldsymbol{X}$. Interesuje nas zmierzenie wpływu na zmienną zależną zmiennych zawartych w macierzy $\boldsymbol{X}_2$. Niech $\widehat{\boldsymbol{\beta}}$ oznacza estymator wektora $\boldsymbol{\beta}$ w modelu z macierzą $\boldsymbol{X}$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

natomiast niech $\widehat{\boldsymbol{\beta}}_1$ będzie estymatorem najmniejszych kwadratów wektora $\boldsymbol{\beta}$ w modelu z ograniczeniami $\boldsymbol{\beta}_2 = 0$ (Zieliński 2007):

$$\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}[\boldsymbol{A}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}]^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}}$$

Tutaj $\boldsymbol{A} = \left[\boldsymbol{0}_{n\times(p-q)} \vdots \boldsymbol{I}_q\right]$ jest macierzą ograniczeń, tzn. $\boldsymbol{\beta}_2 = 0$ jest równoważne temu, że $\boldsymbol{A\beta} = \boldsymbol{0}$ ($\boldsymbol{0}$ oznacza, w zależności od kontekstu, wektor lub macierz zerową, $\boldsymbol{I}_q$ oznacza macierz jednostkową wymiaru q).

Wektor obserwacji $\boldsymbol{Y}$ możemy zapisać jako

$$\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right) + \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right)$$

Zauważmy, że wektory $\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1$ i $\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1$ są ortogonalne. A zatem

$$\left\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2=\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2+\left\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2$$

gdzie $\|\cdot\|^2$ oznacza kwadrat długości wektora (w normie euklidesowej).

Jako miarę wpływu zmiennych zawartych w macierzy $\boldsymbol{X}_2$ przyjmujemy

$$\Re^2(\mathrm{X}_2)=\frac{\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2}{\left\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2}$$

*Twierdzenie***.** Zmienna losowa

$$\frac{\Re^2(\mathbf{X}_2)}{1-\Re^2(\mathbf{X}_2)}\cdot\frac{n-q}{q}$$

ma niecentralny rozkład F z $(q,n-q)$ stopniami swobody i parametrem niecentralności $\boldsymbol{\beta}'\boldsymbol{A}[\boldsymbol{A}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}]^{-1}\boldsymbol{A}\boldsymbol{\beta}$, gdzie $\boldsymbol{A}=\left[\mathbf{0}_{n\times(p-q)}\ \vdots\ \boldsymbol{I}_q\right]$.

*Dowód***.** W dowodzie korzystamy z dobrze znanych faktów dotyczących rozkładów prawdopodobieństwa form liniowych i kwadratowych wektorów losowych o wielowymiarowym rozkładzie normalnym.

Łatwo sprawdzić, że

$$\frac{\Re^2(\boldsymbol{X}_2)}{1-\Re^2(\boldsymbol{X}_2)}=\frac{\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2}{\left\|\boldsymbol{Y}-(\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1)\right\|^2}=\frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']\boldsymbol{A}\widehat{\boldsymbol{\beta}}}{\boldsymbol{Y}'(\boldsymbol{I}_n-\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']\boldsymbol{A})\boldsymbol{Y}}$$

Macierz $\boldsymbol{I}_n-\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']\boldsymbol{A}$ jest idempotentna, więc w mianowniku mamy zmienną losową o centralnym rozkładzie chi-kwadrat z $(n-q)$ stopniami swobody. Ponieważ wektor losowy $\widehat{\boldsymbol{\beta}}$ ma rozkład $N_p(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, więc w liczniku mamy zmienną losową o niecentralnym rozkładzie chi-kwadrat z q stopniami swobody i parametrem niecentralności $\boldsymbol{\beta}'\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']\boldsymbol{A}\boldsymbol{\beta}$. Licznik i Mianownik są niezależnymi zmiennymi losowymi i stąd wynika teza twierdzenia.

Pokażemy teraz, że współczynnik determinacji oraz współczynnik korelacji cząstkowej są szczególnym przypadkami miary $\Re^2(\mathbf{X}_2)$.

Współczynnik determinacji jest miarą wpływu wszystkich zmiennych niezależnych $x_1,\dots,x_m$ na zmienną Y w modelu

$$Y_i=\beta_0+\beta_1x_{1i}+\cdots+\beta_mx_{mi}+\varepsilon_i,\quad i=1,\dots,n.$$

W zapisie macierzowym

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{12} & \cdots & x_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}.$$

Macierz $\boldsymbol{X}$ i wektor parametrów można podzielić na $\boldsymbol{X}_1 = \mathbf{1}_n$,

$$\boldsymbol{X}_2 = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ x_{12} & \cdots & x_{m2} \\ \vdots & \cdots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}, \qquad \boldsymbol{\beta}_1 = [\beta_0], \qquad \boldsymbol{\beta}_2 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}.$$

W tym modelu mamy $p = m + 1$, $q = m$ oraz

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \mathbf{1}_n'\boldsymbol{X}_2 \\ \mathbf{1}_n\boldsymbol{X}_2' & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}, \quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} M & \boldsymbol{k}' \\ \boldsymbol{k} & \boldsymbol{L} \end{bmatrix}$$

gdzie $\boldsymbol{L}^{-1} = \boldsymbol{X}_2'\,\boldsymbol{X}_2 - \frac{1}{n}\boldsymbol{X}_2'\mathbf{1}_n\mathbf{1}_n'\boldsymbol{X}_2$, $\boldsymbol{k}' = -\frac{1}{n}\mathbf{1}_n'\boldsymbol{X}_2\boldsymbol{L}$ oraz $M = \frac{1}{n} + \boldsymbol{k}'\boldsymbol{L}\boldsymbol{k}$ (ogólny wzór na odwrotność blokowej macierzy 2×2 można znaleźć w np. Rao 1982, Zieliński 2007). Estymator najmniejszych kwadratów wektora $\boldsymbol{\beta}$ w pełnym modelu ma postać

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} M & \boldsymbol{k}' \\ \boldsymbol{k} & \boldsymbol{L} \end{bmatrix} \begin{bmatrix} \mathbf{1}'_n \\ \boldsymbol{X}'_2 \end{bmatrix} \boldsymbol{Y} = \begin{bmatrix} M\mathbf{1}'_n + \boldsymbol{k}'\boldsymbol{X}'_2 \\ \boldsymbol{k}\mathbf{1}'_n + \boldsymbol{L}\boldsymbol{X}'_2 \end{bmatrix} \boldsymbol{Y}$$

tzn.

$$\hat{b}_0 = (M\mathbf{1}_n' + \boldsymbol{k}'\boldsymbol{X}_2')\boldsymbol{Y}, \quad \widehat{\boldsymbol{\beta}}_2 = (\boldsymbol{k}\mathbf{1}_n' + \boldsymbol{L}\boldsymbol{X}_2')\boldsymbol{Y}$$

Estymator wektora $\boldsymbol{\beta}_1$ w modelu z ograniczeniami określonymi macierzą $\boldsymbol{A} = [\mathbf{0}_m \vdots \boldsymbol{I}_m]$ wyraża się wzorem

$$\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{I}_{m+1} - \begin{bmatrix} M & \boldsymbol{k}' \\ \boldsymbol{k} & \boldsymbol{L} \end{bmatrix} \begin{bmatrix} \mathbf{0}'_m \\ \boldsymbol{I}_m \end{bmatrix} \left([\mathbf{0}_m \quad \boldsymbol{I}_m] \begin{bmatrix} M & \boldsymbol{k}' \\ \boldsymbol{k} & \boldsymbol{L} \end{bmatrix} \begin{bmatrix} \mathbf{0}'_m \\ \boldsymbol{I}_m \end{bmatrix} \right)^{-1} [\mathbf{0}_m \quad \boldsymbol{I}_m] \right) \widehat{\boldsymbol{\beta}}$$

$$= \begin{bmatrix} 1 & -\boldsymbol{k}'\boldsymbol{L}^{-1} \\ \mathbf{0}_m & \mathbf{0}_{m\times m} \end{bmatrix} \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{b}_0 - \boldsymbol{k}'\boldsymbol{L}^{-1}\widehat{\boldsymbol{\beta}}_2 \\ \mathbf{0}_m \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\,\mathbf{1}'_n\boldsymbol{Y} \\ \mathbf{0}_m \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \mathbf{0}_m \end{bmatrix}$$

gdzie $\bar{Y} = \frac{1}{n}\sum Y_i$. Ponieważ $\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1 = Y\mathbf{1}_n$, więc ($\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$)

$$\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2 = (\widehat{\boldsymbol{Y}} - Y\,\mathbf{1}_n)'(\widehat{\boldsymbol{Y}} - Y\,\mathbf{1}_n) = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

oraz

$$\left\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1\right\|^2=(\boldsymbol{Y}-\bar{Y}\,\mathbf{1}_{\mathrm{n}})'(\boldsymbol{Y}-\bar{Y}\,\mathbf{1}_{\mathrm{n}})=\sum_{\mathrm{i}=1}^{\mathrm{n}}(Y_{\mathrm{i}}-\bar{\mathrm{Y}})^2$$

Współczynnik determinacji, wyrażany zwyczajowo w procentach, określony jest więc w następujący sposób

$$D=\Re^2(\boldsymbol{X}_2)\cdot 100\%=\frac{\Sigma\left(\hat{Y}_i-\bar{Y}\right)^2}{\Sigma(Y_i-\bar{Y})^2}$$

Z twierdzenia otrzymujemy

*Wniosek***.** Zmienna losowa

$$\frac{D}{100-D}\cdot\frac{n-m}{m}$$

ma niecentralny rozkład F z $(m,n-m)$ stopniami swobody i parametrem niecentralności $\boldsymbol{\beta}_2'\boldsymbol{X}_2'(\boldsymbol{I}_n-\mathbf{1}_n\mathbf{1}_n')\boldsymbol{X}_2\boldsymbol{\beta}_2$.

Współczynnik korelacji cząstkowej jest miarą wpływu jednej ze zmiennych niezależnych na zmienną Y w modelu

$$Y_i=\beta_0+\beta_1x_{1i}+\cdots+\beta_mx_{mi}+\varepsilon_i,\quad i=1,\dots,n.$$

Wyprowadzimy odpowiedni wzór dla zmiennej x_1. Dla uproszczenia zapisu model zapisujemy w postaci

$$Y_i=\beta_0+\beta_2x_{2i}+\cdots+\beta_mx_{mi}+\beta_1x_{1i}+\varepsilon_i,\quad i=1,\dots,n.$$

W zapisie macierzowym

$$\boldsymbol{X}=\begin{bmatrix}1 & x_{21} & \cdots & x_{m1} & x_{11}\\ 1 & x_{22} & \cdots & x_{m2} & x_{12}\\ \vdots & \vdots & \cdots & \vdots & \vdots\\ 1 & x_{2n} & \cdots & x_{mn} & x_{1n}\end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \beta_2\\ \vdots\\ \beta_m\\ \beta_1\end{bmatrix}.$$

Macierz $\boldsymbol{X}$ i wektor parametrów można podzielić na

$$\boldsymbol{X}_1=\begin{bmatrix}1 & x_{21} & \cdots & \mathrm{x}_{m1}\\ 1 & x_{22} & \cdots & x_{m2}\\ \vdots & \vdots & \cdots & \vdots\\ 1 & x_{2n} & \cdots & x_{mn}\end{bmatrix}\ \boldsymbol{x}_2=\begin{bmatrix}x_{11}\\ x_{12}\\ \vdots\\ x_{1n}\end{bmatrix}\ \boldsymbol{\beta}_1=\begin{bmatrix}\beta_0\\ \beta_2\\ \vdots\\ \beta_m\end{bmatrix}\quad \boldsymbol{\beta}=[\beta_1]$$

W tym modelu

$$\boldsymbol{X}'\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}'_1\boldsymbol{X}_1 & \boldsymbol{X}'_1\boldsymbol{x}_2\\ \boldsymbol{x}'_2\boldsymbol{X}_1 & \boldsymbol{x}'_2\boldsymbol{x}_2\end{bmatrix}\quad (\boldsymbol{X}'\boldsymbol{X})^{-1}=\begin{bmatrix}\boldsymbol{M} & \boldsymbol{k}\\ \boldsymbol{k}' & L\end{bmatrix}$$

gdzie

$$L = \frac{1}{\boldsymbol{x}_2'\boldsymbol{x}_2 - \boldsymbol{x}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{x}_2},$$

$$\boldsymbol{k}' = -\boldsymbol{L}(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{x}_2\boldsymbol{X}_1, \qquad \boldsymbol{M} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} - L\boldsymbol{k}\boldsymbol{k}'.$$

Ograniczenia opisane są za pomocą wektora $[\mathbf{0}_m' \vdots 1]$. Po wykonaniu odpowiednich obliczeń otrzymujemy, że miarą wpływu zmiennej x_1 na Y jest

$$\Re^2(x_2) = \frac{(\boldsymbol{x}_2'(\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1')\boldsymbol{Y})^2}{(\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1')\boldsymbol{Y})(\boldsymbol{x}_2'(\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1')\boldsymbol{x}_2)}$$

Współczynnik korelacji cząstkowej między Y a x_1 określony jest jako $\sqrt{\Re^2(x_2)}$ i oznaczany $R_{Y(x_1)(x_2,\dots,x_m)}$. Znak tego współczynnika jest oczywiście zgodny ze znakiem $\boldsymbol{x}_2'(\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1')\boldsymbol{Y}$. Z twierdzenia otrzymujemy następujący wniosek

Wniosek. Zmienna losowa

$$\frac{R^2_{Y(x_1)(x_2,\dots,x_m)}}{1 - R^2_{Y(x_1)(x_2,\dots,x_m)}} \cdot (n - m - 1)$$

ma niecentralny rozkład F z $(1, n - m - 1)$ stopniami swobody i parametrem niecentralności $\beta_2^2\boldsymbol{x}_2'(\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1')\boldsymbol{x}_2$.

Dodatkowa zmienna. Niech $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. Przypuśćmy, że do modelu włączamy jeszcze jedną zmienną niezależną, tzn. $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{z}\gamma + \boldsymbol{\eta} = \boldsymbol{W\delta} + \boldsymbol{\eta}$, gdzie $\boldsymbol{W} = [\boldsymbol{X} \vdots \boldsymbol{z}]$, $\boldsymbol{\delta} = [\boldsymbol{\beta} \vdots \gamma]'$ oraz $\boldsymbol{\eta}$ jest wektorem błędów losowych. Tutaj $\boldsymbol{z}$ jest wektorem n-wymiarowym reprezentującym włączaną zmienną, natomiast γ jest nieznanym współczynnikiem regresji. Wyznaczymy współczynnik $\Re^2(\boldsymbol{W})$ i porównamy go z $\Re^2(\boldsymbol{X})$.

Estymatorem najmniejszych kwadratów wektora $\boldsymbol{\delta}$ jest oczywiście $\widehat{\boldsymbol{\delta}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{Y}$. Mamy

$$\boldsymbol{W}'\boldsymbol{W} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{z} \\ \boldsymbol{z}'\boldsymbol{X} & \boldsymbol{z}'\boldsymbol{z} \end{bmatrix}$$

oraz

$$(\boldsymbol{W}'\boldsymbol{W})^{-1} = \frac{1}{c}\begin{bmatrix} c(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{a}\boldsymbol{a}' & -\boldsymbol{a} \\ -\boldsymbol{a}' & 1 \end{bmatrix}$$

gdzie $\boldsymbol{a} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{z}$ oraz $c = \boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{z}$. Zatem

$$\boldsymbol{d} = \begin{bmatrix} \boldsymbol{b} - \dfrac{\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}}{\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{z}}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z} \\ \dfrac{\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}}{\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{z}} \end{bmatrix}$$

Ponieważ

$$\boldsymbol{Wd} = \boldsymbol{Xb} + \frac{\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}}{\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{z}}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{z}$$

więc

$$\|\boldsymbol{Wd}\|^2 = \|\boldsymbol{Xb}\|^2 + \frac{(\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y})^2}{\boldsymbol{z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{z}}$$

Zatem $\Re^2(\boldsymbol{W}) \geq \Re^2(\boldsymbol{X})$, tzn. model z dodatkową zmienną nie gorzej odtwarza zmienną zależną niż model bez niej.

LITERATURA

Borkowski B., Dudek H., Szczesny W. 2003: Ekonometria, wybrane zagadnienia, PWN, Warszawa.
Draper N. R., Smith H. 1973: Analiza regresji stosowana, PWN, Warszawa.
Rao C. R. 1982: Modele liniowe statystyki matematycznej, PWN, Warszawa.
Seber G. A. F. 1977: Linear Regression Analysis, Wiley, New York.
Zieliński W. 2007: Teoretyczne podstawy ekonometrycznych jednorównaniowych modeli liniowych, Wydawnictwo SGGW, Warszawa.

Generalized measure of fit in a linear model

Abstract: In a model of linear regression two measures of fit are used: coefficient of determination and coefficient of partial correlation. A generalization of those measures is proposed and a measure of influence of a group of independent variables excluding an influence of remaining variables is constructed.

Keywords: measure of fit, determination coefficient, partial correlation, linear regression