
PROSTA REGRESJA LINIOWA II

Jako estymator wariancji błędu w modelu prostej regresji liniowej przyjmujemy

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Nieobciążoność estymatora s^2 dla prostej liniowej regresji można pokazać w następujących krokach.

$$y_i \hat{y}_i = y_i - b_0 - b_1 x_i = y_i - (\bar{y} - b_1 \bar{x}) - b_1 x_i = (y_i - \bar{y}) - b_1 (x_i - \bar{x})$$

Tak więc

$$\sum_{i=1}^n (y_i - \bar{y}_i) = \sum_{i=1}^n (y_i - \bar{y}) - b_1 \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Zauważ, że $(y_i - \bar{y}_i)x_i = [(y_i - \bar{y}) - b_1(x_i - \bar{x})]x_i$, stąd mamy

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)x_i &= \sum_{i=1}^n [(y_i - \bar{y}) - b_1(x_i - \bar{x})]x_i \\ &= \sum_{i=1}^n [(y_i - \bar{y}) - b_1(x_i - \bar{x})](x_i - \bar{x}) \\ &= \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) - b_1 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= n(S_{xy} - b_1 S_{xx}) = n\left(S_{xy} - \frac{S_{xy}}{S_{xx}} S_{xx}\right) = 0 \end{aligned}$$

Aby pokazać, że s^2 jest nieobciążonym estymatorem wariancji błędu, najpierw zauważamy, że

$$(y_i - \hat{y}_i)^2 = [(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2,$$

w związku z tym,

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)^2 &= \sum_{i=1}^n [(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2nb_1 S_{xy} + nb_1^2 S_{xx} \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2n \frac{S_{xy}}{S_{xx}} S_{xy} + n \frac{S_{xy}^2}{S_{xx}^2} S_{xx} \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - n \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

Ponieważ

$$(y_i - \bar{y})^2 = [\beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})]^2$$

i

$$(y_i - \bar{y})^2 = \beta_1^2(x_i - \bar{x})^2 + (\epsilon_i - \bar{\epsilon})^2 + 2\beta_1(x_i - \bar{x})(\epsilon_i - \bar{\epsilon}),$$

stąd

$$E(y_i - \bar{y})^2 = \beta_1^2 E(x_i - \bar{x})^2 + E(\epsilon_i - \bar{\epsilon})^2 + 2\beta_1 E(x_i - \bar{x})(\epsilon_i - \bar{\epsilon}),$$

i

$$\sum_{i=1}^n E(y_i - \bar{y})^2 = n\beta_1^2 S_{xx} + \sum_{i=1}^n \frac{n-1}{n} \sigma^2 = n\beta_1^2 S_{xx} + (n-1)\sigma^2.$$

Co więcej mamy

$$\begin{aligned} E(S_{xy}) &= E\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right) \\ &= \frac{1}{n} E \sum_{i=1}^n (x_i - \bar{x}) y_i \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) E y_i \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i) \\ &= \frac{1}{n} \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i \\ &= \frac{1}{n} \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 = \beta_1 S_{xx} \end{aligned}$$

i

$$\text{Var}(S_{xy}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) y_i\right) = \frac{1}{n} \sum_{i=1}^n \text{Var}(y_i) = \frac{1}{n} S_{xx} \sigma^2.$$

Tak więc możemy napisać

$$E(S_{xy}^2) = \text{Var}(S_{xy}) + [E(S_{xy})]^2 = \frac{1}{n} S_{xx} \sigma^2 + \beta_1^2 S_{xx}^2,$$

i

$$E\left(\frac{n S_{xy}^2}{S_{xx}}\right) = \sigma^2 + n \beta_1^2 S_{xx}.$$

Ostatecznie, $E(\hat{\sigma}^2)$ jest równe:

$$E \sum_{i=1}^n (y_i - \hat{y})^2 = n \beta_1^2 S_{xx} + (n-1)\sigma^2 - n \beta_1^2 S_{xx} - \sigma^2 = (n-2)\sigma^2.$$

Innymi słowami udowodniliśmy, że

$$E(s^2) = E\left(\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2\right) = \sigma^2.$$

Zatem s^2 , jako estymator wariancji błędu, jest nieobciążonym estymatorem wariancji błędu σ^2 w prostej regresji liniowej.

1 Forma macierzowa regresji liniowej z wieloma zmiennymi objaśniającymi.

W wielu badaniach naukowych często konieczne jest ustalenie relacji między zmienną odpowiedzi (zależną) (y) a więcej niż jedną zmienną objaśniającą (zmienną niezależną). Oznaczając przez $(x_1, x_2, \dots, x_k)$ wektor zmiennych objaśniających ogólna forma modelu wielowymiarowej regresji liniowej na postać:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon,$$

gdzie ϵ jest błędem losowym. W tym przypadku regresory $(x_1, x_2, \dots, x_k)$ mogą zawierać regresory i ich potęgi. W klasycznej postaci zakłada się, że składnik błędu ϵ ma rozkład normalny ze średnią 0 i stałą wariancją σ^2 . W pierwszym momencie możemy przez analogię z prostą regresji liniową interpretować postać regresji wielowymiarowej z hiperpłaszczyzną odpowiedzi. Jednak niektóre regresory mogą mieć postać innych regresorów podniesionych do wyższych potęg, lub mogą nawet być funkcjami regresorów, o ile funkcje te nie zawierają nieznanymi parametrów. Zatem model regresji wielowymiarowej może mieć postać powierzchni odpowiedzi o różnorodnych kształtach. Następująca definicja liniowej regresji wielowymiarowej pozwala zrozumieć różnicę między modelem liniowym a modelem nieliniowym.

Definicja. Model liniowy jest definiowany jako model, w którym występuje liniowa zależność od parametrów β_i .

Tak więc poniżej przedstawione przykłady definiuje modele wielowymiarowej regresji liniowej, w którym zmienna zależna y jest funkcją liniową parametrów regresji:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \epsilon, \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3 + \epsilon, \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 \log(x_1) + \beta_3 \log(x_3) + \epsilon. \end{aligned}$$

Natomiast następujące przykłady definiują modele nieliniowe:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2^{\beta_3} + \epsilon, \\ y &= \frac{1}{\lambda + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3)} + \epsilon. \end{aligned}$$

Definicja. Niech $\mathbf{y}^T = (y_1, y_2, \dots, y_n)$ będzie wektorem, a a_{ij} gdzie $i, j = 1, 2, \dots, n$ tablicą $n \times n$ liczb rzeczywistych.

Forma kwadratowa zmiennych $y_1, y_2, \dots, y_n$ jest zdefiniowana jako

$$f(y_1, y_2, \dots, y_n) = \sum_{i,j=1}^n a_{ij} y_i y_j.$$

Formę kwadratową można zapisać w postaci macierzowej: $\mathbf{y}^T \mathbf{A} \mathbf{y}$ gdzie $\mathbf{A} = (a_{ij})_{n \times n}$ jest macierzą rozmiaru $n \times n$. Formy kwadratowe odgrywają ważną rolę w analizie modelu wielowymiarowej regresji liniowej. W klasycznym podejściu parametry modelu regresji liniowej są szacowane poprzez minimalizację sumy kwadratów reszt:

$$\mathbf{b} = \arg \min_{(\beta_0, \beta_1, \dots, \beta_k)} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki})]^2,$$

gdzie $\mathbf{b} = (b_1, b_2, \dots, b_k)$. Tak więc kwadrat reszt jest w rzeczywistości formą kwadratową.

1.1 Estymacja metodą najmniejszych kwadratów parametrów wielowymiarowej regresji liniowej.

Model wielowymiarowej regresji liniowej jest zazwyczaj zapisywany w następującej postaci:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i,$$

gdzie y_i jest zmienną zależną, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ to współczynniki regresji, a ϵ_i to błędy losowe, o których zakładamy, że $E(\epsilon_i) = 0$ i $\text{Var}(\epsilon_i) = \sigma^2$ dla $i = 1, 2, \dots, n$. W klasycznym modelu regresji zakłada się, że błędy mają rozkład normalny z wspólną wartością wariancji σ^2 . Współczynniki regresji są szacowane przy użyciu metody najmniejszych kwadratów. Należy zauważyć, że nie trzeba zakładać, że rozkład błędów regresji jest zgodny z rozkładem normalnym aby dokonać metodą najmniejszych kwadratów estymacji współczynników regresji. Stosunkowo łatwo pokazać, że przy założeniu normalności rozkładu błędów estymacja współczynników regresji metodą najmniejszych kwadratów daje dokładnie taką samą postać estymatorów jak metoda największej wiarygodności. Postać modelu wielowymiarowej regresji liniowej można wyrazić w formie macierzowej:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

gdzie

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{pmatrix}.$$

Postać macierzowa modelu regresji pozwala przeanalizować i przedstawić wiele właściwości modelu regresji wygodniej i bardziej efektywnie. Zauważmy, że prosta regresja liniowa jest szczególnym przypadkiem wielowymiarowej regresji liniowej i może być przedstawiona w formie macierzowej. Estymacja metodą najmniejszych kwadratów polega na rozwiązaniu następującego zagadnienia optymalizacyjnego:

$$\mathbf{b} = \arg \min_{\boldsymbol{\beta}} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})],$$

gdzie $\mathbf{b} = (b_1, b_2, \dots, b_k)$, jest k -wymiarowym wektorem wartości wyestymowanych współczynników regresji.

Twierdzenie. Estymatorem wyznaczonym metodą najmniejszych kwadratów dla współczynników modelu wielowymiarowej regresji liniowej $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ mają przy założeniu, że $(\mathbf{X}^T \mathbf{X})$ jest niesingularną macierzą następującą postać

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Niesingularność macierzy $(\mathbf{X}^T \mathbf{X})$ jest równoważna założeniu, że kolumny macierzy $\mathbf{X}$ są niezależnymi liniowo wektorami.

Dowód. W celu wyznaczenia estymatorów metodą najmniejszych kwadratów musimy zminimalizować sumę kwadratów reszt co jest równoważne rozwiązaniu następującego równania:

$$\frac{\partial}{\partial \mathbf{b}} [(\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})] = 0,$$

lub równoważnego mu równania postaci:

$$\frac{\partial}{\partial \mathbf{b}}[(\mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X} \mathbf{y} + \mathbf{y}^T \mathbf{X}^T \mathbf{X} \mathbf{b})] = 0.$$

Różniczkując względem każdej współrzędnej wektora β - otrzymujemy następujące równanie normalne modelu wielowymiarowej regresji liniowej:

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{y}.$$

Ponieważ macierz $\mathbf{X}^T \mathbf{X}$ jest niesingularna otrzymujemy, że $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, co kończy dowód.

Omówimy teraz własności statystyczne estymatorów najmniejszych kwadratów współczynników regresji. Najpierw omawiamy nieobciążoność estymatora najmniejszych kwadratów współczynnika $\mathbf{b}$.

Twierdzenie. *Estymator $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ jest nieobciążonym estymatorem β . Dodatkowo*

$$\text{Var}(\mathbf{b}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2.$$

Dowód. Zauważmy, że

$$\mathbf{E} \mathbf{b} = \mathbf{E}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{E}(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta.$$

Co kończy dowód nieobciążoności estymatora $\mathbf{b}$. Wariancję estymatora β możemy obliczyć w następujący sposób:

$$\begin{aligned} \text{Var}(\mathbf{b}) &= \text{Var}((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}) \\ &= (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \text{Var}(\mathbf{b}) (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \\ &= (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 = (\mathbf{X}' \mathbf{X})^{-1} \sigma^2. \end{aligned}$$

Kolejnym parametrem w klasycznym modelu regresji liniowej jest wariancja σ^2 , wielkość, która jest nieobserwowana. Wnioskowanie statystyczne na temat współczynników regresji oraz diagnostyka modelu regresji silnie zależy od estymacji wariancji błędu σ^2 . Aby oszacować σ^2 , rozważymy sumę kwadratów reszt:

$$\mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X} \mathbf{b})^T (\mathbf{y} - \mathbf{X} \mathbf{b}) = \mathbf{y}^T [\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T)] \mathbf{y} = \mathbf{y}^T \mathbf{P} \mathbf{y}.$$

To jest odległość między wektorem obserwacji $\mathbf{y}$ i dopasowanego metodą regresji wektora $\hat{\mathbf{y}}$. Macierz $\mathbf{P} = [\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T)]$ jest idempotentna, czyli

$$\mathbf{P}^2 = [\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T)] [\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T)] + \mathbf{P}$$

Stąd wiemy, że wartości własne macierzy $\mathbf{P}$ są równe 0 lub 1. Zauważ, że macierz $\mathbf{X} (\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T)$ jest również idempotentna. Tak więc, mamy

$$\begin{aligned} \text{rank}(\mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}') &= \text{tr}(\mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}') \\ &= \text{tr}(\mathbf{X}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1}) = \text{tr}(\mathbf{I}_p) = p. \end{aligned}$$

Ponieważ $tr(A - B) = tr(A) - tr(B)$ więc

$$\begin{aligned} rank(I - X(X'X)^{-1}X') &= tr(I - X(X'X)^{-1}X') \\ &= tr(I_n) - tr(X'X(X'X)^{-1}) = n - p \end{aligned}$$

Suma kwadratów reszt w wielowymiarowej regresji liniowej $\mathbf{e}^T \mathbf{e}$ możemy zapisać jako formę kwadratową wektora odpowiedzi $\mathbf{y}$.

$$\mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}^T (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y}.$$

Biorąc wartość oczekiwaną otrzymamy Powyższe rozważania możemy podsumować w postaci

$$\begin{aligned} E(\mathbf{e}'\mathbf{e}) &= E[\mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y}] \\ &= (X\beta)'(I - X(X'X)^{-1}X')(X\beta) + \sigma^2(n - p) \\ &= (X\beta)'(X\beta - X(X'X)^{-1}X'X\beta) + \sigma^2(n - p) = \sigma^2(n - p) \end{aligned}$$

następującego twierdzenia.

Twierdzenie. Nieobciążony estymator wariancji w modelu wielowymiarowej regresji liniowej jest postaci

$$s^2 = \frac{\mathbf{e}^T \mathbf{e}}{n - p} = \frac{\mathbf{y}^T (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y}}{n - p} = \frac{1}{n - p} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Niech $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ Wektor $\mathbf{y}$ można podzielić na dwa wektory

$$(\mathbf{I} - \mathbf{P})\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y} \quad \text{ i } \quad \mathbf{P}\mathbf{y} = (\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y}.$$

Zakładając normalność błędu regresji otrzymujemy niezależność składników rozkładu $(\mathbf{I} - \mathbf{P})\mathbf{y}$ i $\mathbf{P}\mathbf{y}$. Aby to pokazać, po prostu obliczymy kowariancję $(\mathbf{I} - \mathbf{P})\mathbf{y}$ i $\mathbf{P}\mathbf{y}$.

$$\begin{aligned} &\text{Cov}((I - P)\mathbf{y}, P\mathbf{y}) \\ &= (I - P)\text{Cov}(\mathbf{y})P = (I - P)P\sigma^2 \\ &= (I - X(X'X)^{-1}X')X(X'X)^{-1}X'\sigma^2 \\ &= [X(X'X)^{-1}X' - (X(X'X)^{-1}X')X(X'X)^{-1}X']\sigma^2 \\ &= (X(X'X)^{-1}X' - X(X'X)^{-1}X')\sigma^2 = 0 \end{aligned}$$

Ponieważ $(\mathbf{I} - \mathbf{P})\mathbf{y}$ i $\mathbf{P}\mathbf{y}$ są wektorami normalnymi, zerowa kowariancja implikuje ich niezależność, a zatem również $\mathbf{y}^T (\mathbf{I} - \mathbf{P})\mathbf{y}$ i $\mathbf{y}^T \mathbf{P}\mathbf{y}$ są również niezależne. Gdy $\mathbf{P}$ jest idempotentna to funkcja kwadratowa wektora o rozkładzie normalnym $\mathbf{y}^T \mathbf{P}\mathbf{y}$ ma rozkład chi-kwadrat z $p = rank(\mathbf{P})$ stopniami swobody. Ta własność pozwala skonstruować testu F , który jest powszechnie stosowana w problemach testowania hipotez dla wielowymiarowej regresji liniowej.

Twierdzenie. Niech $\mathbf{y}$ ma rozkład $N(\mu, I)$ oraz niech A i B będą dwiema macierzami. Wektory o rozkładzie normalnym $A\mathbf{y}$ i $B\mathbf{y}$ są niezależne wtedy i tylko wtedy, gdy $AB^T = 0$.

Dowód. Przypomnijmy, że niezależność dwóch normalnych wektorów jest równoważna zerowej kowariancji między nimi. Obliczamy kowariancję $A\mathbf{y}$ i $B\mathbf{y}$.

$$\text{Cov}(\mathbf{A}\mathbf{y}, \mathbf{B}\mathbf{y}) = \mathbf{A} \text{Cov}(\mathbf{y}) \mathbf{B}^T = \mathbf{A}\mathbf{B}^T.$$

Zatem niezależność wektorów $\mathbf{A}\mathbf{y}$ i $\mathbf{B}\mathbf{y}$ jest równoważna warunkowi $\mathbf{A}\mathbf{B}^T = 0$.

Korzystając z tego twierdzenia, możemy łatwo wykazać, że $(\mathbf{I} - \mathbf{P})\mathbf{y}$ i $\mathbf{P}\mathbf{y}$ są niezależne. Ponieważ $\mathbf{P}$ jest idempotentna, zatem

$$(\mathbf{I} - \mathbf{P})\mathbf{P} = \mathbf{P} - \mathbf{P}^2 = \mathbf{P} - \mathbf{P} = 0.$$

1.2 Forma macierzowa prostej regresji liniowej

Model prostej regresji liniowej jest szczególnym przypadkiem wielowymiarowego modelu liniowych i może być wyrażony w formie macierzowej. W szczególności,

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{pmatrix}.$$

Dla postaci macierzowej wzór na $\mathbf{b}$ w przypadku jednowymiarowym mamy

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

Macierz odwrotną dla macierzy $\mathbf{X}'\mathbf{X}$ ma postać:

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \\ &= \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}. \end{aligned}$$

Estymator wyznaczony metodą najmniejszych kwadratów w przypadku jednowymiarowym w postaci macierzowej ma postać:

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\ &= \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n \sum x_i y_i - \sum x_i \sum y_i \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix}. \end{aligned}$$

Stąd otrzymujemy

$$\mathbf{b}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\sum (x_i - \bar{x})^2} = \bar{y} - \mathbf{b}_1 \bar{x}$$

oraz

$$\mathbf{b}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

Krzysztof Topolski