

Regresja liniowa

Komentarz do wykładu z 3. kwietnia

Dla zmiennej losowej (X, Y) znamy niezależne wartości $(x_1, y_1), \dots, (x_n, y_n)$. Można też przyjąć, że rozważamy n niezależnych zmiennych $(X_1, Y_1), \dots, (X_n, Y_n)$ o tym samym rozkładzie każda. Szukamy liczb β_0, β_1 minimalizujących wartość funkcji

$$f(\beta_0, \beta_1) = \sum_{i=1}^n (\beta_0 + \beta_1 x_i - y_i)^2. \quad (1)$$

Innymi słowy, chcemy znaleźć prostą o równaniu $y = \beta_0 + \beta_1 x$, dla której suma odległości (kwadratów odległości) od punktów (x_i, y_i) jest minimalna. Odległość jest tutaj rozumiana jako odległość punktu od jego rzutu pionowego na prostą.

Metoda analityczna

Obliczmy pochodne cząstkowe funkcji $f(\beta_0, \beta_1)$ względem zmiennych β_0, β_1 . Jest

$$\begin{cases} \frac{\partial f}{\partial \beta_0} = 2 \cdot \sum_{i=1}^n (\beta_0 + \beta_1 x_i - y_i) = 0, \\ \frac{\partial f}{\partial \beta_1} = 2 \cdot \sum_{i=1}^n (\beta_0 + \beta_1 x_i - y_i) \cdot x_i = 0. \end{cases}$$

Po uporządkowaniu otrzymamy równania

$$\begin{cases} n \cdot \beta_0 + n \cdot \bar{x} \beta_1 = n \cdot \bar{y}, \\ n \cdot \bar{x} \beta_0 + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases} \quad (2)$$

Z pierwszego z równań (2) znajdujemy zależność $\beta_0 = \bar{y} - \beta_1 \bar{x}$, i po podstawieniu do równania drugiego $\beta_1 = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2}$. Korzystając z zadania 1.7 stwierdzamy, że $\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Zmienna (X, Y) ma rozkład określony następująco:

$\begin{matrix} X \\ Y \end{matrix}$	x_1	x_2	\dots	x_n
y_1	$1/n$	0	\dots	0
y_2	0	$1/n$	\dots	0
\vdots	\vdots	\vdots	\ddots	\vdots
y_n	0	0	\dots	0

. Przy takim in-

terpretowaniu obserwacji otrzymujemy wzory

$$\beta_0 = \bar{y} - \beta_1 \bar{x}, \quad \beta_1 = \frac{\text{Cov}(X, Y)}{V(X)}. \quad (3)$$

Definicja 1. Prostą o równaniu $Y = \beta_0 + \beta_1 \cdot X$, gdzie β_0, β_1 określone są wzorem (3), nazywamy prostą regresji zmiennej Y względem zmiennej X .

Uwaga:

W dotychczasowych rozważaniach znaleźliśmy ekstremum funkcji f we wzorze (1). Intuicyjnie: szukamy ekstremum wielowymiarowej paraboli o ramionach skierowanych ku górze. Formalnie, powinniśmy sprawdzić, że

$$\frac{\partial^2 f}{\partial \beta_0^2} > 0, \quad \begin{vmatrix} \frac{\partial^2 f}{\partial \beta_0^2} & \frac{\partial^2 f}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 f}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 f}{\partial \beta_1^2} \end{vmatrix} > 0.$$

Obliczając pochodne cząstkowe drugiego rzędu otrzymujemy nierówności

$$\frac{\partial^2 f}{\partial \beta_0^2} = n > 0, \quad \left| \begin{array}{cc} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{array} \right| = n \cdot \sum x_i^2 - n^2 \bar{x}^2 = n \sum (x_i - \bar{x})^2 > 0.$$

Regresja zapisana macierzowo.

Znalezienie współczynników regresji polega na wyznaczeniu współczynników β_0, β_1 takich, że

$$y_i \approx \beta_0 + \beta_1 x_i, \quad (i = 1, \dots, n). \quad (4)$$

Macierzowo, można równania (4) zapisać w postaci

$$\begin{array}{ccc} X & \beta & \approx Y \\ \left[\begin{array}{cc} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{array} \right] & \left[\begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right] & \approx \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right]. \end{array} \quad (5)$$

Na ogół układ równań (5) możemy rozwiązać tylko w sposób przybliżony. Wynika to z postawienia zadania regresji. Szukamy prostej położonej **blisko** punktów (x_k, y_k) , nie istnieje prosta przechodząca **przez** te punkty (chyba, że wszystkie y_k to jakaś wielokrotność x_k).

Po pomnożeniu równości (5) prawostronnie przez macierz X^T otrzymujemy zależność w której można postawić już znak $=$, a mianowicie $X^T X \cdot \beta = X^T Y$. Przy założeniu, że macierz $X^T X$ jest nieosobliwa daje to wzór

$$\beta = (X^T X)^{-1} X^T Y. \quad (6)$$

Jeżeli $\text{rk}(X^T X) = 1$, to stąd wynika, że $\text{rk}(X) = 1$.¹ Dla wartości x_i jest zatem $x_1 = \dots = x_n$, czyli można pominąć zmienną X , a więc poszukujemy zależności o postaci $Y = \beta_0$. Drugi ze wzorów (3) miałby postać $\frac{0}{0}$.

Aproksymacja średniokwadratowa.

Elementami macierzy $X^T X$ są iloczyny skalarne wierszy macierzy X^T i kolumn macierzy X , czyli iloczyny skalarne kolumn macierzy X (to znaczy $\{1, X\}$). Układ równań (2) można zatem przepisać w postaci macierzowej jako

$$\left[\begin{array}{cc} \langle 1, 1 \rangle & \langle 1, X \rangle \\ \langle X, 1 \rangle & \langle X, X \rangle \end{array} \right] \cdot \left[\begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right] = \left[\begin{array}{c} \langle 1, Y \rangle \\ \langle X, Y \rangle \end{array} \right]. \quad (7)$$

Ostatni wzór to współczynniki elementu najlepszej aproksymacji dla funkcji $Y(X)$ z podprzestrzeni rozpiętej przez $\{1, X\}$.² Układ równań (7) to układ równań normalnych.

Regresja (względem) wielu zmiennych.

ZADANIE: Dane są niezależne obserwacje $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$, dla $i = 1, \dots, n$. Znaleźć wektor $\beta = [\beta_0, \beta_1, \dots, \beta_k]$ minimalizujący wartość funkcji

$$f(\beta) = \sum_{i=1}^n (\beta_0 + \beta_1 x_i - y_i)^2. \quad (8)$$

¹Notatki z algebry – lemat 4.20

²Analiza numeryczna: n -ty wielomian optymalny dla funkcji $f(x)$ to wielomian z przestrzeni rozpiętej przez funkcje $\{x^0, x^1, \dots, x^n\}$

W podejściu analitycznym należy rozwiązać układ równań

$$\frac{\partial f}{\partial \beta_0} = \frac{\partial f}{\partial \beta_1} = \dots = \frac{\partial f}{\partial \beta_k} = 0,$$

niezbyt skomplikowany, ale trudno poddający się zwięzłej notacji. Przejdźmy zatem do notacji macierzowej, podobnej do wzoru (5).

$$\begin{array}{c} X \\ \left[\begin{array}{ccccc} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{array} \right] \end{array} \begin{array}{c} \beta \\ \left[\begin{array}{c} \beta_0 \\ \vdots \\ \beta_k \end{array} \right] \end{array} \approx \begin{array}{c} Y \\ \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right] \end{array}. \quad (9)$$

Mnożąc powyższą równość, lewostronnie przez X^T otrzymujemy **równość** $X^T X \beta = X^T Y$, identyczną z równością (6). Podobnie też, – jak poprzednio – spełnienie warunku $\text{rk}(X^T X) < k + 1$ oznacza iż [któraś/któreś] ze zmiennych X_1, \dots, X_k [jest/są] [zbędna/zbędne], jako kombinacja liniowa pozostałych.³

Nawiązując do aproksymacji średniokwadratowej – szukamy elementu najlepszej aproksymacji dla funkcji Y z przestrzeni (podprzestrzeni “regularnych” funkcji) rozpiętej przez funkcje $\{1, X_1, \dots, X_k\}$.

Podsumowanie w jednym zdaniu: $\beta = (X^T X)^{-1} X^T Y$.

Regresja kwadratowa i wielomianowa.

ZADANIE: Dane są niezależne obserwacje (x_i, y_i) , dla $i = 1, \dots, n$. Znaleźć wektor $\beta = [\beta_0, \beta_1, \beta_2]^T$ minimalizujący wartość funkcji

$$f(\beta) = \sum_{i=1}^n (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 - y_i)^2. \quad (10)$$

Sposób analityczny to rozwiązanie układu trzech równań z trzema niewiadomymi $\beta_1, \beta_1, \beta_2$:

$$\frac{\partial f}{\partial \beta_0} = \frac{\partial f}{\partial \beta_1} = \frac{\partial f}{\partial \beta_2} = 0.$$

Prostszym sposobem jest notacja macierzowa. Zamiast zmiennej X wprowadźmy dwie zmienne $X_1 = X$, $X_2 = X^2$. Oznacza to iż mamy do czynienia z zadaniem regresji zmiennej Y względem zmiennych X_1, X_2 . Niech $x_{i1} \leftarrow x_i$ oraz $x_{i2} \leftarrow x_i^2$. Równanie macierzowe (9) można przepisać w postaci

$$\begin{array}{c} X \\ \left[\begin{array}{ccc} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{array} \right] \end{array} \begin{array}{c} \beta \\ \left[\begin{array}{c} \beta_0 \\ \beta_1 \\ \beta_2 \end{array} \right] \end{array} \approx \begin{array}{c} Y \\ \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right] \end{array}. \quad (11)$$

Stąd, ponownie, podsumowanie w jednym zdaniu: $\beta = (X^T X)^{-1} X^T Y$.

Zadanie regresji wielomianowej jest określone następująco:

ZADANIE: Dane są niezależne obserwacje (x_i, y_i) , dla $i = 1, \dots, n$. Znaleźć wektor $\beta = [\beta_0, \dots, \beta_k]^T$

³Dodanie nawiasów $[]$ oznacza utworzenie wektora, algebra raz jeszcze. $\left[\begin{array}{c} \text{któraś} \\ \text{któreś} \end{array} \right] \sim \left[\begin{array}{c} \text{jest} \\ \text{są} \end{array} \right] + \left[\begin{array}{c} \text{zbędna} \\ \text{zbędne} \end{array} \right]$

minimalizujący wartość funkcji

$$f(\beta) = \sum_{i=1}^n (\beta_0 + \beta_1 x_i + \dots + \beta_k x_i^k - y_i)^2. \quad (12)$$

Równanie macierzowe (9) ponownie okazuje się najbardziej intuicyjne. Jest mianowicie

$$\begin{matrix} & X & & \\ \left[\begin{array}{ccccc} 1 & x_1 & x_1^2 & \dots & x_1^k \\ 1 & x_2 & x_2^2 & \dots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^k \end{array} \right] & \begin{matrix} \beta \\ \beta_0 \\ \vdots \\ \beta_k \end{matrix} & \approx & \begin{matrix} Y \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix} \end{matrix}. \quad (13)$$

Ponownie, po raz kolejny, podsumowanie w jednym zdaniu: $\beta = (X^T X)^{-1} X^T Y$. Zauważmy również, że oprócz potęg wartości x_k można dołączyć do równania regresji zmienne $\log(x_k)$, $\exp(x_k)$, $\sin(x_k)$ i różne inne przekształcenia x_k . W efekcie – regresje: wielomianowa, logarytmiczna, wykładnicza, trygonometryczna i nie_wiadomo_jaka sprowadzają się do regresji wielu zmiennych.

←

Z poważaniem,
Witold Karczewski