Regresja liniowa

Komentarz do wykładu z 19. kwietnia

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

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

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

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

Po uporządkowaniu otrzymamy równania

$$\begin{cases} n \cdot \beta_0 + n \cdot \bar{x}\, \beta_1 = n \cdot \bar{y}, \\ n \cdot \bar{x}\, \beta_0 + \beta_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{2}$$

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

Zmienna (X, Y) ma rozkład określony następująco:

$Y \backslash X$	x_1	x_2	$\ldots$	x_n
y_1	$1/n$	0	$\ldots$	0
y_2	0	$1/n$	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	0	0	$\ldots$	$1/n$

. Przy takim interpretowaniu obserwacji otrzymujemy wzory

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

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

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

Obliczając pochodne cząstkowe drugiego rzędu otrzymujemy nierówności

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

Regresja zapisana macierzowo.

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

$$y_i \approx \beta_0 + \beta_1\, x_i, \quad (i = 1, \ldots, n). \tag{4}$$

Macierzowo, można równania (4) zapisać w postaci

$$\begin{array}{ccc} X & \beta & \approx & Y \\ \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} & \approx & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{array}. \tag{5}$$

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 $\underset{2\times 2}{X^T X} \cdot \underset{2\times 1}{\beta} = \underset{2\times 1}{X^T Y}$. Przy założeniu, że macierz $X^T X$ jest nieosobliwa daje to wzór

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

Jeżeli $\mathrm{rk}(X^T X) = 1$, to stąd wynika, że $\mathrm{rk}(X) = 1$.[1] Dla wartości x_i jest zatem $x_1 = \ldots = x_n$, 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 $\{\mathbb{1}, X\}$). Układ równań (2) można zatem przepisać w postaci macierzowej jako

$$\begin{bmatrix} \langle \mathbb{1}, \mathbb{1} \rangle & \langle \mathbb{1}, X \rangle \\ \langle X, \mathbb{1} \rangle & \langle X, X \rangle \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \langle \mathbb{1}, Y \rangle \\ \langle X, Y \rangle \end{bmatrix}. \tag{7}$$

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

Regresja (względem) wielu zmiennych.

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

$$f(\beta) = \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k - y_i\right)^2. \tag{8}$$

[1]Notatki z algebry – lemat 4.20

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

W podejściu analitycznym należy rozwiązać układ równań

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

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

$$\begin{array}{cccc} X & \beta & \approx & Y \\ \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} & \approx & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{array}. \tag{9}$$

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 $\mathrm{rk}(X^T X) < k+1$ oznacza iż [którąś/któreś] ze zmiennych $X_1, \ldots, X_k$ [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 $\{\mathbb{1}, X_1, \ldots, X_k\}$.

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

Regresja kwadratowa i wielomianowa.

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

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

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

$$\begin{array}{cccc} X & \beta & \approx & Y \\ \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} & \approx & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{array}. \tag{11}$$

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

Zadanie regresji wielomianowej jest określone następująco:
Zadanie: Dane są niezależne obserwacje (x_i, y_i), dla $i = 1, \ldots, n$. Znaleźć wektor $\beta = [\beta_0, \ldots, \beta_k]^T$

[3] Dodanie nawiasów [] oznacza utworzenie wektora, algebra raz jeszcze. $\begin{bmatrix} \text{którąś} \\ \text{któreś} \end{bmatrix} \sim \begin{bmatrix} \text{jest} \\ \text{są} \end{bmatrix} + \begin{bmatrix} \text{zbędna} \\ \text{zbędne} \end{bmatrix}$

minimalizujący wartość funkcji

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

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

$$\begin{array}{cccc} X & \beta & \approx & Y \\ \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^k \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} & \approx & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{array} . \tag{13}$$

Ponownie, po raz kolejny, podsumowanie w jednym zdaniu: $\beta = \left(X^T X\right)^{-1} X^T Y$. 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.

$\mapsfrom$

Z poważaniem,
Witold Karczewski