Регрессионный анализ

Этимология: «регрессия» – отступление, возврат

x – отклонение от среднего роста отца
y – отклонение от среднего роста сына
$\Big\} \Leftarrow$ Положительная связь, но тенденция возврата (отклонение у сына < отклонения у отца)

Закономерность – функция регрессии.

Классическая линейная модель множественной регрессии (КЛММР)

$y_i = \theta_0 + \theta_1 x_i^{(1)} + ... + \theta_p x_i^{(p)} + \varepsilon_i$, $i = 1,...,n$.

1) $M\varepsilon_i = 0$, $i = 1,...,n$;

2) $M(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}$ – гомоскедастичность; – взаимная некоррелированность;

3) $rank\, X = p + 1 < n$ – одна объясняющая переменная не выражается через другую, существует $\left(X^T X\right)^{-1}$, если $p + 1 \geq n$, то для выводов недостаточно данных.

Матричная форма:

$$Y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_1^{(0)} = 1 & x_1^{(1)} & ... & x_1^{(p)} \\ & ... & & \\ x_n^{(0)} = 1 & x_n^{(1)} & ... & x_n^{(p)} \end{pmatrix} \quad \Theta = \begin{pmatrix} \theta_0 \\ ... \\ \theta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{pmatrix} \quad \sum\nolimits_\varepsilon = \begin{pmatrix} M(\varepsilon_1^2) & ... & M(\varepsilon_1 \varepsilon_n) \\ & ... & \\ M(\varepsilon_n \varepsilon_1) & ... & M(\varepsilon_n^2) \end{pmatrix}.$$

$$Y = X\Theta + \varepsilon, \quad M\varepsilon = \mathbf{0_n}, \quad \sum\nolimits_\varepsilon = \sigma^2 E_n, \quad rank\, X = p + 1 < n.$$

↑ ковариационная матрица остатков

Если $\varepsilon \sim N\left(0; \sigma^2 E_n\right)$, то нормальная КЛММР.

Оценивание неизвестных параметров. Метод наименьших квадратов (МНК)

Модельные значения $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i^{(1)} + ... + \hat{\theta}_p x_i^{(p)}$ минимально отличаются от наблюдаемых y_i:

$$\sum_{i=1}^{n} \varepsilon_i^2 \to \min_{\theta_0,...,\theta_p}, \quad \varepsilon_i = y_i - \theta_0 - \theta_1 x_i^{(1)} - ... - \theta_p x_i^{(p)}.$$

Матричная форма:

$$\varepsilon = Y - X\Theta, \quad (Y - X\Theta)^T (Y - X\Theta) \to \min_\Theta, \qquad \boxed{(AB)^T = B^T A^T}$$

$$Y^T Y - 2\Theta^T X^T Y + \Theta^T X^T X\Theta \to \min_\Theta, \quad -2X^T Y + 2X^T X\Theta = 0,$$

$$\boxed{\hat{\Theta} = \left(X^T X\right)^{-1} X^T Y}$$

← невырожденная

Случай парной регрессии:

$$X = \begin{pmatrix} 1 & x_1 \\ ... & \\ 1 & x_p \end{pmatrix} \quad X^T X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \quad X^T Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

$$\begin{cases} n\theta_0 + \theta_1 \sum x_i = \sum y_i, \\ \theta_0 \sum x_i + \theta_1 \sum x_i^2 = \sum x_i y_i. \end{cases} \qquad \theta_0 = \frac{\sum y_i - \theta_1 \sum x_i}{n},$$

$$\sum x_i \sum y_i - \theta_1 \left(\sum x_i\right)^2 + \theta_1 n \sum x_i^2 = n \sum x_i y_i, \qquad \boxed{\hat{\theta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}}, \quad \boxed{\hat{\theta}_0 = \bar{y} - \theta_1 \bar{x}}$$

Помесячные данные о печати фотографий в некоторой фирме

	месяц	y, кол-во, шт.	$z^{(1)}$, цена, руб.	$z^{(2)}$, рекл., руб.	$z^{(3)}$, праздники	$z^{(i)}$, индекс цен
1	январь, 2003	12 500	2,5	0	3	1
2	февраль	7 600	3	0	1	0,99
3	март	6 900	3	0	1	1,01
4	апрель	13 500	3	5 000	0	1,01
5	май	9 700	3	0	3	1,03
6	июнь	10 700	3	2 000	1	1,04
7	июль	12 100	3	2 000	0	1,05
8	август	9 700	3,5	2 000	0	1,03
9	сентябрь	7 000	4	2 000	0	1,05
10	октябрь	7 200	4	2 000	0	1,05
11	ноябрь	8 200	4	2 000	1	1,06
12	декабрь	8 400	4	2 000	1	1,1
13	январь, 2004	13 100	4	2 000	3	1,11
14	февраль	8 700	4	0	1	1,12
15	март	12 200	4	5 000	1	1,14
16	апрель	6 900	4	0	0	1,16
17	май	6 200	4	0	3	1,17
18	июнь	9 600	4	0	1	1,19
19	июль	8 700	4	0	0	1,18
20	август	11 900	4	4 000	0	1,18
21	сентябрь	12 600	4	6 000	0	1,2
22	октябрь	7 900	4	1 000	0	1,22
23	ноябрь	9 300	4	2 000	1	1,24
24	декабрь	11 800	4	2 000	1	1,27

	y	$x^{(1)} = z^{(1)} / z^{(i)}$	$x^{(2)} = z^{(2)} / z^{(i)}$	$x^{(3)} = z^{(3)}$	$\hat{y}$, прогноз	ε, остатки
1	12 500	2,50	0	3	11 552	948
2	7 600	3,03	0	1	8 792	–1 192
3	6 900	2,97	0	1	8 968	–2 068
4	13 500	2,97	4 950	0	13 975	–475
5	9 700	2,91	0	3	10 339	–639
6	10 700	2,88	1 923	1	11 398	–698
7	12 100	2,86	1 905	0	10 857	1 243
8	9 700	3,40	1 942	0	9 310	390
9	7 000	3,81	1 905	0	8 059	–1 059
10	7 200	3,81	1 905	0	8 059	–859
11	8 200	3,77	1 887	1	8 745	–545
12	8 400	3,64	1 818	1	9 071	–671
13	13 100	3,60	1 802	3	10 350	2 750
14	8 700	3,57	0	1	7 202	1 498
15	12 200	3,51	4 386	1	12 354	–154
16	6 900	3,45	0	0	6 963	–63
17	6 200	3,42	0	3	8 852	–2 652
18	9 600	3,36	0	1	7 819	1 781
19	8 700	3,39	0	0	7 135	1 565
20	11 900	3,39	3 390	0	10 974	926
21	12 600	3,33	5 000	0	12 964	–364
22	7 900	3,28	820	0	8 390	–490
23	9 300	3,23	1 613	1	10 044	–744
24	11 800	3,15	1 575	1	10 225	1 575

$$X^T X = \begin{pmatrix} 24 & 79{,}23 & 36820 & 22 \\ 79{,}23 & 264{,}18 & 123426 & 70{,}42 \\ 36820 & 123426 & 114453587 & 18607 \\ 22 & 70{,}42 & 18607 & 46 \end{pmatrix}, \quad \left(X^T X\right)^{-1} = \begin{pmatrix} 4{,}83 & -1{,}39 & -0{,}00003 & -0{,}17 \\ -1{,}39 & 0{,}414 & -0{,}000005 & 0{,}0326 \\ -0{,}00003 & -0{,}000005 & 0{,}00000002 & 0{,}00001 \\ -0{,}17 & 0{,}0326 & 0{,}00001 & 0{,}0483 \end{pmatrix},$$

$$X^T Y = \begin{pmatrix} 232400 \\ 760351 \\ 407592667 \\ 217900 \end{pmatrix}, \quad \hat{\Theta} = \begin{pmatrix} 17094 \\ -2938 \\ 1{,}1327 \\ 600{,}86 \end{pmatrix}, \qquad \hat{y}_i = 17094 - 2938 x_i^{(1)} + 1{,}1327 x_i^{(2)} + 600{,}86 x_i^{(3)}.$$