Линейная регрессия

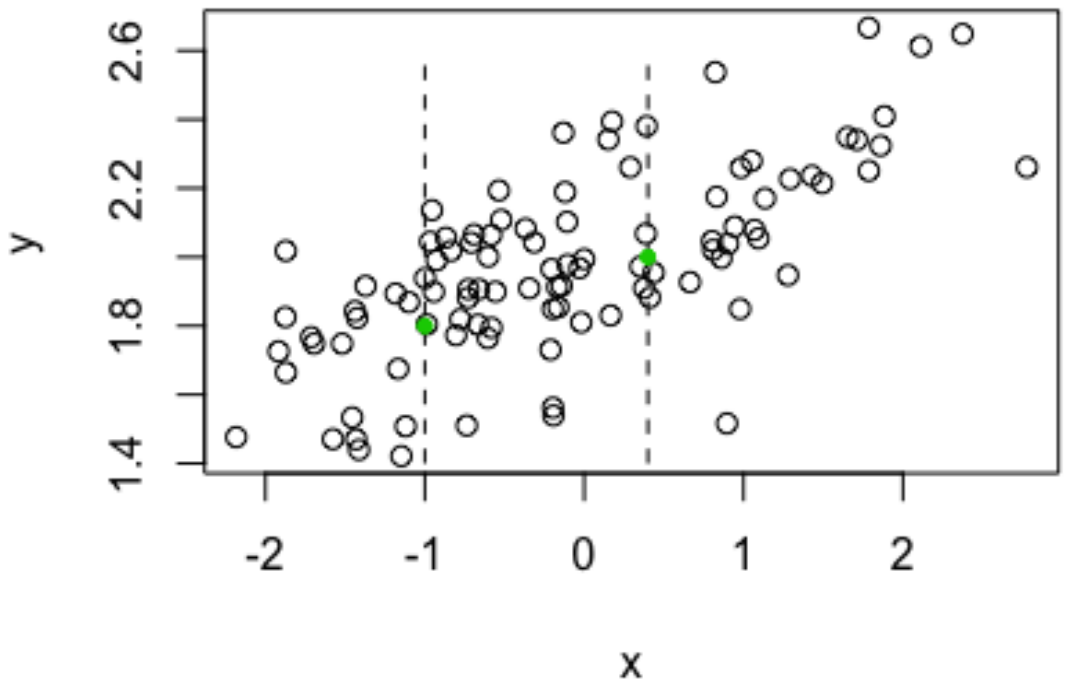
Н.П. Алексеева,

СПбГУ, математико-механический факультет

2020 г.

Кривая регрессии

Кривую регрессии для η по переменной ξ определяют как траекторию точки $(x, E(\eta|\xi = x))$. При $g(\xi) = E(\eta|\xi = x)$ достигается минимум выражения $\mathbb{E}(\eta - g(\xi))^2$.



Статистическая модель линейной регрессии

Если в качестве функций $g(\xi)$ рассматривать выражения вида $\alpha + \beta\xi$, то прямая, при которой достигается минимум, называется линейной средней квадратической регрессией.

Пусть (x_i, y_i) двумерная выборка наблюдений случайных величин ξ и η, $i = 1, \ldots, n$, и

$$y_i = \alpha + \beta x_i + \delta_i \tag{1}$$

где ошибки δ_i такие, что $E\delta_i = 0, \quad D\delta_i = \sigma^2, \quad E\delta_i\delta_j = 0$. Из всего семейства прямых вида $g(x) = \alpha + \beta x$ выбирается „наилучшая“ в том смысле, что квадратичная форма

$$Q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

принимает наименьшее значение.

Оценка параметров линейной регрессии

Минимизируем квадратичную форму

$$Q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

$$\begin{cases} -\frac{1}{2}\frac{\partial Q}{\partial \alpha} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \\ -\frac{1}{2}\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0 \end{cases} \iff$$

$$\iff \begin{cases} \sum_{i=1}^{n} y_i - \alpha n - \beta \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i x_i - \alpha \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

Учитывая то, что $\sum_{i=1}^{n} x_i = n\bar{x}$ и $\sum_{i=1}^{n} y_i = n\bar{y}$, получаем

$$\alpha = \bar{y} - \beta \bar{x}\,.$$

Подставляем $\alpha = \bar{y} - \beta\bar{x}$ во второе уравнение

$$\sum_{i=1}^{n} y_i x_i - \alpha \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - (\bar{y} - \beta\bar{x}) n\bar{x} - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\beta = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Напомним, что

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) =$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Пример оценки параметров

$(x_i, y_i), i = 1, 2, \ldots, n$, парная выборка массы тела до и после диеты, $\bar{x} = 74.7, \bar{y} = 72.1$, $\sum_{i=1}^{n} x_i^2 = 56655$, $\sum_{i=1}^{n} y_i^2 = 52805$, $\sum_{i=1}^{n} x_i y_i = 54599$, $R = 0.88$.

	1	2	3	4	5	6	7	8	9	10
x_i	68	80	92	81	70	79	78	66	57	76
y_i	60	84	87	79	74	71	72	67	57	70
$x_i y_i$	4080	6720	8004	6399	5180	5609	5616	4422	3249	5320
x_i^2	4624	6400	8464	6561	4900	6241	6084	4356	3249	5776
y_i^2	3600	7056	7569	6241	5476	5041	5184	4489	3249	4900

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{54599 - 10 \cdot 74.7 \cdot 72.1}{56655 - 10 \cdot 74.7^2} = 0.8667,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 72.1 - 0.8667 \cdot 74.7 = 7.353\,.$$

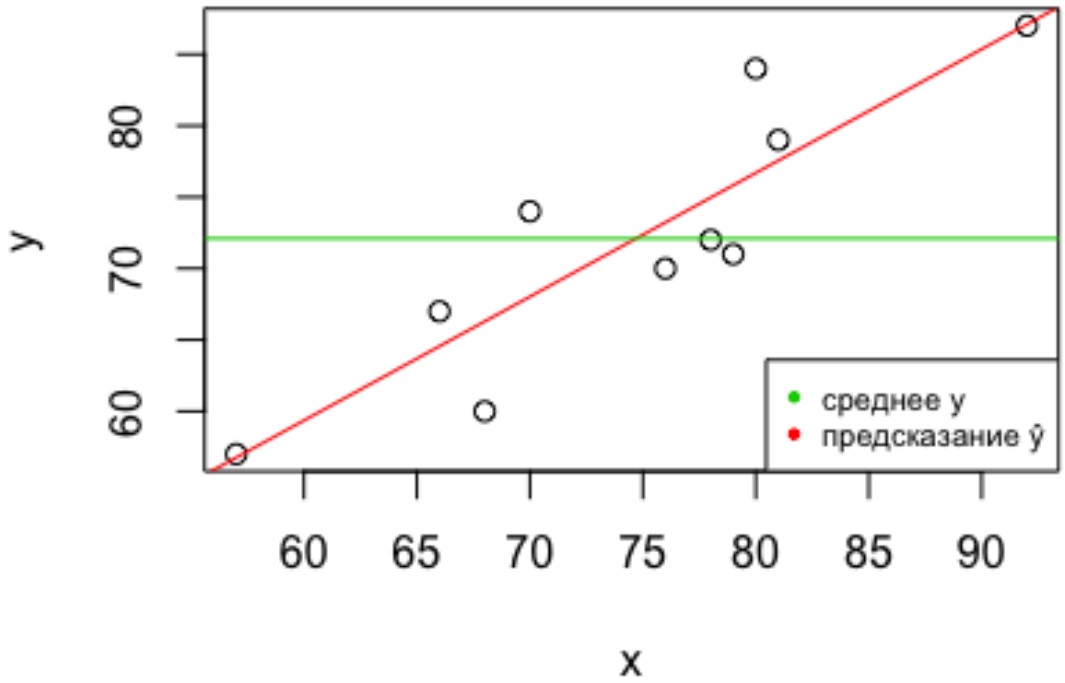


Рис.: $y = 0.8667x + 7.353$.

Некоррелированность остатка с независимой переменной

Разность $\tilde{y}_i = y_i - \hat{\beta}x_i - \hat{\alpha}$ называется остатком.

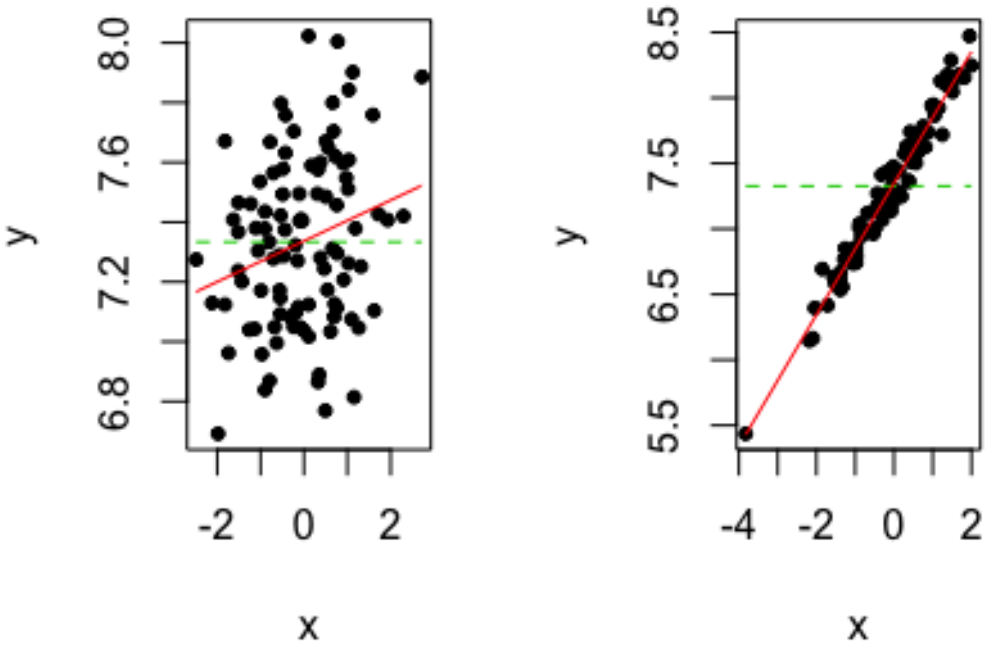
$\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i = 0$

Действительно,
$\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}x_i - \hat{\alpha}) = \bar{y} - \hat{\beta}\bar{x} - \hat{\alpha} = 0$, так как
$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

$\sum_{i=1}^{n}(\tilde{y}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i)(x_i - \bar{x}) = 0$

$\sum_{i=1}^{n}\tilde{y}_i(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \hat{\beta}x_i - \hat{\alpha})(x_i - \bar{x}) =$
$= \sum_{i=1}^{n}(y_i - \hat{\beta}x_i - \bar{y} + \hat{\beta}\bar{x})(x_i - \bar{x}) =$
$= \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$ по определению $\hat{\beta}$.

Источники вариации



$$Q_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \qquad Q_E = \sum_{j=1}^{n} (y_j - \hat{y}_i)^2$$

Источники вариации: Q_R обусловлен отклонением регрессии от среднего, Q_E – отклонением наблюдений от линии регрессии.

Источники вариации

$$Q = \sum_{i=1}^{n}(y_i - \bar{y})^2, Q_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, Q_E = \sum_{j=1}^{n}(y_j - \hat{y}_i)^2 \implies Q = Q_R + Q_E.$$

$$Q = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 =$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Так как $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, то $\hat{y}_i - \bar{y} = \hat{\beta}(x_i - \bar{x})$, и
$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) =$

$$= \sum_{i=1}^{n}\hat{\beta}(y_i - \hat{y}_i)(x_i - \bar{x}) = \hat{\beta}\sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))(x_i - \bar{x}) =$$

$$= \hat{\beta}\left(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = 0$$

из определения $\hat{\beta} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Коэффициент детерминации

Коэффициент детерминации вводится как

$$R^2 = 1 - \frac{Q_E}{Q}, \tag{2}$$

где $Q = \sum_{i=1}^{n} (y_i - \bar{y})^2$ – общая сумма квадратов отклонений от выборочного среднего, Q_E – источник вариации, обусловленный отклонением наблюдений от линии регрессии.

Пример

j	Годы	Продолж-ть жизни ξ_1	Нац.богатсво (млрд.дол.) ξ_2	Водка (бут.в год) ξ_3
1	1970	68.9	2372*	25.3
2	1975	68.1	2372*	28
3	1980	67.6	2489	30
4	1985	69.2	3379	23.5
5	1990	69.2	4130	18
6	1995	64.6	1171	38.4
7	1998	67	689	29.6

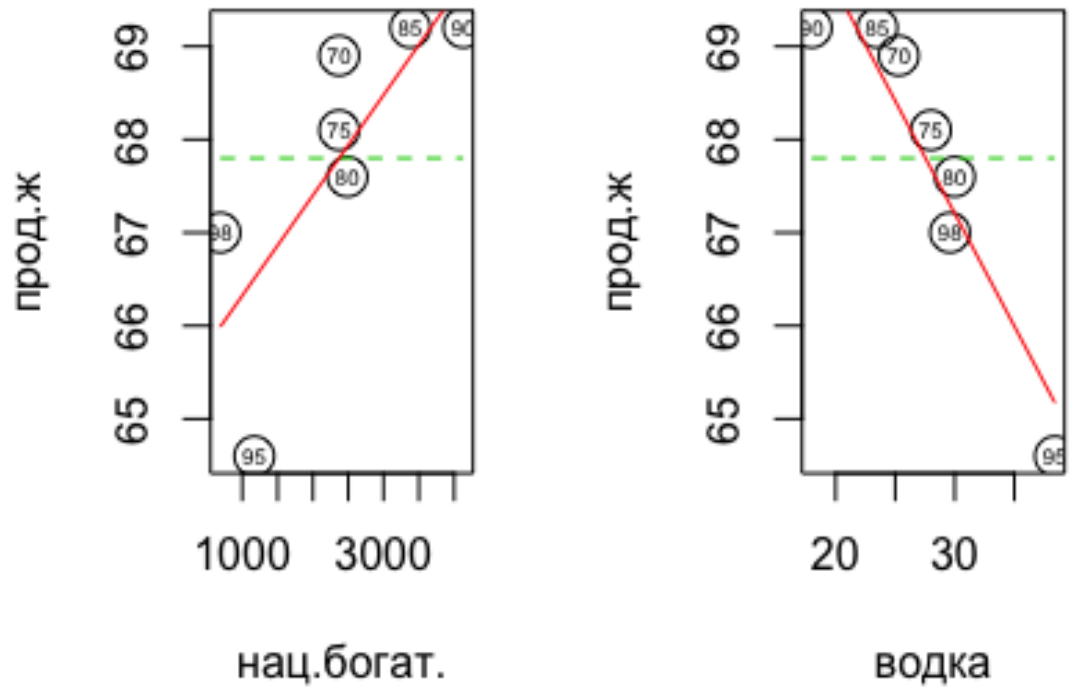


Рис.: $r_1 = 0.776, R_1^2 = 0.6, p = 0.04,$
$r_2 = -0.934, R_2^2 = 0.87, p = 0.002.$

Решение задач 1

1. Производятся 200 испытаний, которые завершаются успехом с вероятностью 0.85. Чему равна вероятность того, что число успехов будет в пределах от 160 до 180? Чему равна вероятность получить значение более 171?

Пусть ξ случайное число успехов в $n = 200$ независимых испытаниях с вероятностью успеха $p = 0.85$, $q = 0.15$. Применяем теорему Муавра-Лапласа, $m_1 = 160$, $m_2 = 180$. $np = 170$, $npq = 25.5$,

$$P\{m_1 < \xi < m_2\} = \Phi\left(\frac{m_2 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{m_1 - np}{\sqrt{npq}}\right) =$$

$$= \Phi(1.98) - \Phi(-1.98) = 2\Phi(1.98) - 1 = 2 \cdot 0.9761 - 1 = 0.9523$$

$$P\{\xi > m_3\} = 1 - \Phi\left(\frac{m_3 - np}{\sqrt{npq}}\right) = 1 - \Phi(0.1980) = 0.4215$$

Решение задачи 2

2. Плотность распределения имеет вид: $f(x) = 0.2x + 0.3$ на интервале от 0 до 2, иначе она равна нулю. Проверить основное свойство плотности и найти математическое ожидание.

$$\int_0^2 (0.2x + 0.3)dx = \frac{0.2x^2}{2} + 0.3x \Big| \quad {}_0^2 = 0.4 + 0.6 = 1$$

$$\int_0^2 x\,(0.2x + 0.3)\,dx = \int_0^2 (0.2x^2 + 0.3x)dx =$$

$$= \frac{0.2x^3}{3} + \frac{0.3x^2}{2} \Big| \quad {}_0^2 = \frac{8}{15} + \frac{3}{5} = 1\frac{2}{15}$$

Решение задачи 3

3. Чему равна дисперсия равномерно распределенной на интервале от 4 до 7 случайной величины?

Плотность равномерно распределенной на интервале от a до b случайной величины ξ имеет вид

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0, & \text{иначе} \end{cases} = \begin{cases} \frac{1}{3} & x \in [4, 7] \\ 0, & \text{иначе} \end{cases}$$

$$\mathbb{E}\xi = \int_4^7 \frac{x}{3} dx = \frac{1}{3}\frac{x^2}{2}|_4^7 = \frac{7^2 - 4^2}{2 \cdot 3} = \frac{11}{2} = 5.5,$$

$$\mathbb{E}\xi^2 = \int_4^7 \frac{x^2}{3} dx = \frac{1}{3}\frac{x^3}{3}|_4^7 = \frac{7^3 - 4^3}{3 \cdot 3} = 31,$$

$$\mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = 31 - \left(\frac{11}{2}\right)^2 = 0.75\,.$$

Решение задачи 4

4. Произведено 100 испытаний, из которых было 79 успехов. Построить 95%-й доверительный интервал для вероятности успеха p.

Пусть $x_1, \ldots, x_n$ бернуллиевские случайные величины с вероятностью успеха p. $\mathbb{E}x_i = p$, $\mathbb{D}\xi = \sigma^2 = p(1-p)$. Оценим параметры по выборке: $\hat{p} = \bar{x} = \frac{79}{100}$, $\hat{\sigma^2} = \bar{x}(1-\bar{x}) = \frac{79 \cdot 21}{100 \cdot 100}$, $\alpha = 0.05$.

$$\bar{x} \sim \mathcal{N}\left(p, \frac{\sigma}{\sqrt{n}}\right) \iff z = \frac{\bar{x} - p}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)\,.$$

$$P\{-z_{1-\frac{\alpha}{2}} < \frac{\bar{x} - p}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\} = 1 - \alpha, \quad P\{p_1 < p \leq p_2\} = 1 - \alpha,$$

$$p_1 = \bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 0.757484, \quad p_2 = \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 0.822515$$