РЕГРЕССИОННЫЙ АНАЛИЗ

ПРИМЕРЫ

1. **Моделирование числа поступивших в университет для лучшего понимания факторов, удерживающих детей в том же учебном заведении.**

2. **Моделирование потоков миграции в зависимости от таких факторов как средний уровень зарплат, наличие медицинских, школьных учреждений, географическое положение…**

3. **Моделирование дорожных аварий как функции скорости, дорожных условий, погоды и т.д.,**

4. **Моделирование потерь от пожаров как функции от таких переменных как количество пожарных станций, время обработки вызова, или цена собственности. Суть регрессионного анализа заключается в нахождении наиболее важных факторов, которые влияют на зависимую переменную.**

статистический метод исследования влияния одной или нескольких независимых переменных $X_1, X_2, \ldots, X_n$ на зависимую переменную Y.

Независимые переменные иначе называют *регрессорами или*

$$Y = f(x_1, x_2, x_3) + \varepsilon,$$

где $f(x_1, x_2, x_3)$ — детерминированная составляющая отклика Y, зависящая от x_1, x_2, x_3, а ε — случайная составляющая.

1) простая линейная регрессия

$$Y = \beta_0 + \beta_1 x + \varepsilon;$$

2) множественная регрессия

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon;$$

3) полиномиальная регрессия

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_{k-1} x^{k-1} + \varepsilon;$$

4) регрессионная модель общего вида:

$$Y = \beta_0 + \beta_1 \varphi_1(x_1, x_2, \ldots, x_m) + \ldots + \beta_{k-1} \varphi_{k-1}(x_1, x_2, \ldots, x_m) + \varepsilon,$$

где $\varphi_i(x_1, x_2, \ldots, x_m)$, $i = 1, 2, \ldots, k-1$, — заданные функции факторов.

Коэффициенты $\beta_0, \beta_1, \ldots, \beta_{k-1}$ называются ***параметрами регрессии***.

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Количественная характеристика степени линейной зависимости между случайными величинами X и Y

Оценка коэффициента корреляции ρ, по выборке наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, вычисляется по формуле

$$r = \frac{Q_{xy}}{\sqrt{Q_x \cdot Q_y}},$$

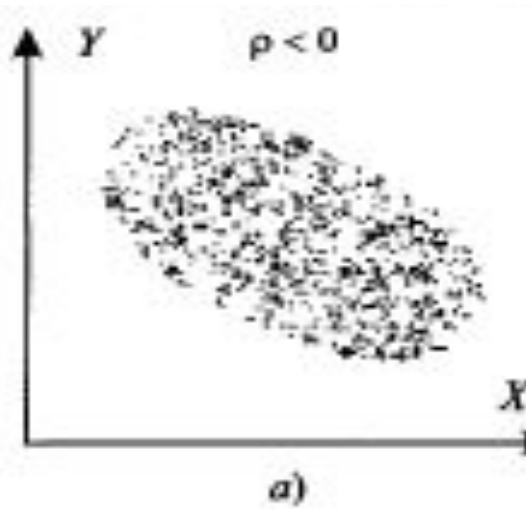
$$Q_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n};$$

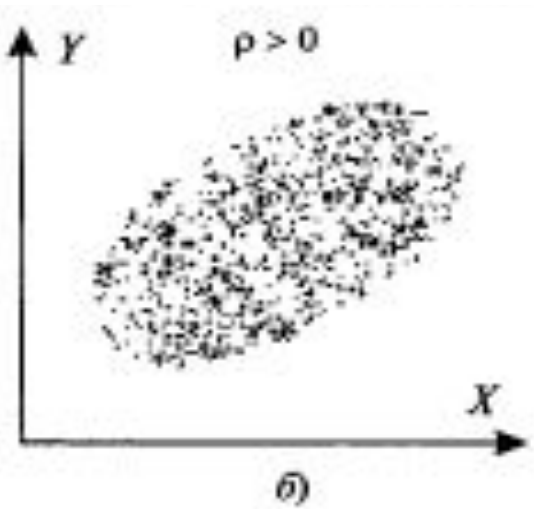
$$Q_y = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n};$$

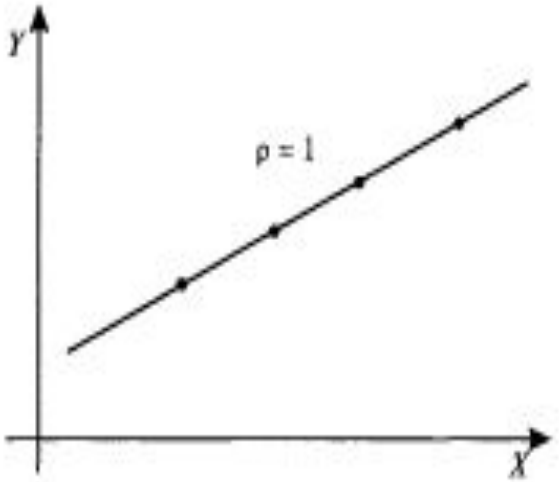
$$Q_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n};$$

$$\bar{x} = \frac{1}{n}\sum x_i, \; \bar{y} = \frac{1}{n}\sum y_i.$$

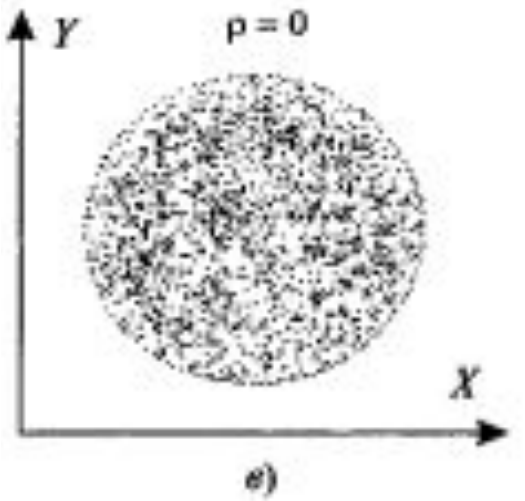
ВАРИАНТЫ РАСПОЛОЖЕНИЯ «ОБЛАКА» ТОЧЕК

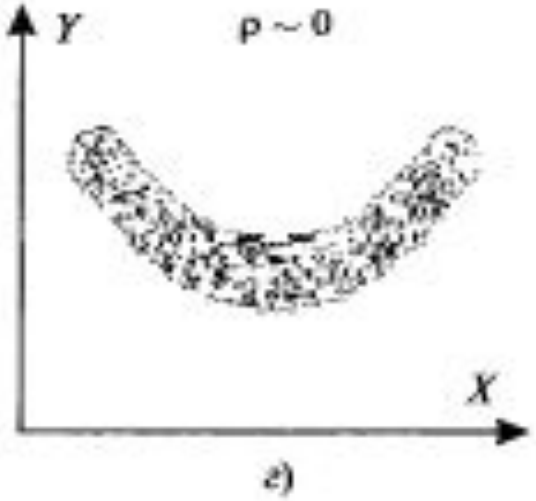






Величины *X* и *Y* *некоррелированы*, т.е. между ними нет линейной зависимости





$$-1 \le \rho \le 1$$

УСЛОВИЕ ПРИМЕНЕНИЯ УРАВНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

когда между случайными величинами X и Y существует достаточно тесная линейная статистическая зависи $|r| > 0$

$$y = \beta_0 + \beta_1 x,$$

где β_0 и β_1 — параметры линейной регрессии; x — независимая переменная (фактор, предиктор); y — зависимая переменная (отклик).

ОЦЕНКА ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ *Y* на *x*

$\overline{\beta_0}$, $\overline{\beta_1}$ } **значения, минимизирующие** $Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$,

$Q(\beta_0, \beta_1)$ сумма квадратов отклонений значений зависимой переменной y_i от значений $\tilde{y}_i$, вычисляемых по уравнению регрессии: $\tilde{y}_i = \beta_0 + \beta_1 x_i$, $i = 1, 2, \ldots, n$,

$$\tilde{\beta}_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = \frac{Q_{xy}}{Q_x}$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}.$$

ОЦЕНКА ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ *X*

$$\tilde{\beta}_1' = \frac{Q_{xy}}{Q_y};$$

$$\tilde{\beta}_0' = \bar{x} - \tilde{\beta}_1' \bar{y}.$$

Уравнения

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$$

и

$$x = \tilde{\beta}_0' + \tilde{\beta}_1' y = \bar{x} + r\frac{s_x}{s_y}(y - \bar{y}),$$

где s_x и s_y — оценки средних квадратических отклонений σ_x и σ_y:

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{Q_x}{n}}, \quad s_y = \sqrt{s_y^2} = \sqrt{\frac{Q_y}{n}},$$

r — оценка коэффициента корреляции ρ, называются *выборочными уравнениями линейной регрессии.*

ЗАДАЧА ЛИНЕЙНОГО РЕГРЕССИОННОГО АНАЛИЗА

по результатам наблюдений $(x_i, y_i), i = 1, 2, \dots, n$

а) получить наилучшие точечные и интервальные оценки неизвестных параметров β_0, β_1 и σ^2;

б) проверить статистические гипотезы о параметрах модели;

в) проверить, достаточно ли хорошо модель согласуется с результатами наблюдений (адекватность модели результатам наблюдений).

. Несмещенная оценка дисперсии ошибок наблюдений S^2 связана с распределением $\chi^2(n-2)$ следующим соотношением

$$\frac{S^2}{\sigma^2} = \frac{\chi^2(n-2)}{n-2}, \quad S^2 = \frac{Q_e}{n-2}.$$

Всюду в дальнейшем будем предполагать, что ошибки наблюдений ε_i, $i = 1, 2, \ldots, n$ имеют нормальное распределение: $\varepsilon_i \sim N(0, \sigma^2)$ и независимы.

Это предположение эквивалентно тому, что результаты наблюдений y_i, $i = 1, 2, \ldots, n$ являются реализациями независимых нормально распределенных случайных величин Y_i:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, 2, \ldots, n.$$

Оценки $\tilde{\beta}_0$ и $\tilde{\beta}_1$ параметров линейной регрессии, вычисленные по формулам (1) и (2), как линейные функции Y_i, $i = 1, 2, \ldots, n$ также будут случайными величинами, имеющими нормальное распределение:

$$\tilde{\beta}_0 \sim N(\beta_0, D[\tilde{\beta}_0]),$$

$$\tilde{\beta}_1 \sim N(\beta_1, D[\tilde{\beta}_1]),$$

где оценки дисперсии $\tilde{\beta}_0$ и $\tilde{\beta}_1$ соответственно равны:

$$D[\tilde{\beta}_0] = \frac{\sigma^2(\sum x_i^2)}{nQ_x}, \quad D[\tilde{\beta}_1] = \frac{\sigma^2}{Q_x}.$$

Доверительный интервал для дисперсии ошибок наблюдений σ^2 имеет вид

$$\frac{(n-2)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-2)} < \sigma^2 < \frac{(n-2)S^2}{\chi^2_{\frac{\alpha}{2}}(n-2)},$$

где $\chi^2_p(n-2)$ — квантили распределения χ^2 с $(n-2)$ степенями свободы порядка p, а S^2 — оценка дисперсии ошибок наблюдений.

МАТРИЧНОЕ ПРЕДСТАВЛЕНИЕ ЗАДАЧИ ЛИНЕЙНОГО РЕГРЕССИОННОГО АНАЛИЗА

регрессионная матрица $(n \times 2)$ $A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, вектор $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$,

вектор параметров модели $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$,

вектор ошибок наблюдений $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$,

Тогда простая линейная регрессия определяется матричным уравнением

$$Y = A\beta + \varepsilon.$$

ЗАДАНИЕ НА ПРАКТИЧЕСКУЮ РАБОТУ

1		2		3		4		5		6		7	
–2,2	–4,0	–0,5	3,3	1,2	2,8	2,8	4,0	1,4	–0,7	–1,0	–1,2	2,7	1,0
–0,1	0,2	0,9	0,5	–3,0	–1,1	0,6	4,1	–2,3	3,9	2,5	2,4	0,2	2,8
3,1	5,4	1,5	0,0	–0,4	3,0	3,0	2,9	0,2	1,6	3,3	5,3	–1,2	2,9
–0,2	0,7	0,6	1,0	2,3	4,3	–1,6	5,9	4,8	–2,7	2,2	3,9	–0,5	3,2
1,0	3,5	–0,2	1,7	–0,3	1,3	–0,7	5,5	1,2	–0,6	2,0	0,5	–0,7	2,5

8		9		10		11		12		13		14	
3,3	3,8	–0,2	4,5	3,5	2,0	4,2	1,5	1,8	7,1	6,2	2,5	5,3	3,5
1,1	3,7	0,8	6,2	1,5	–0,6	5,3	1,8	4,3	7,8	3,1	2,0	2,8	0,7
–1,4	–1,1	–1,2	3,2	3,5	4,7	1,4	2,5	0,0	1,8	1,7	0,6	3,0	2,1
2,7	3,5	–0,5	2,9	1,8	–0,7	0,3	2,9	1,3	2,3	8,0	3,6	3,5	1,3
0,8	2,5	1,0	5,3	–0,3	0,4	2,0	2,8	3,2	4,7	6,1	1,5	3,0	2,5

15		16		17		18		19		20		21	
5,6	1,0	3,7	2,8	9,2	0,5	1,7	2,3	4,8	6,8	4,4	4,7	7,1	5,4
6,1	2,5	3,7	2,5	1,2	7,8	4,9	5,0	3,0	7,8	3,8	6,4	4,5	6,5
6,1	2,2	3,9	1,3	4,3	5,3	0,6	2,5	5,8	6,1	2,4	4,3	8,3	4,6
5,6	5,5	3,9	1,3	6,5	4,5	8,0	7,1	4,5	6,8	3,4	4,3	6,7	5,4
2,7	6,0	1,9	3,4	4,5	4,2	1,0	3,1	6,5	4,5	2,3	2,7	10,4	4,0