Линейная регрессия

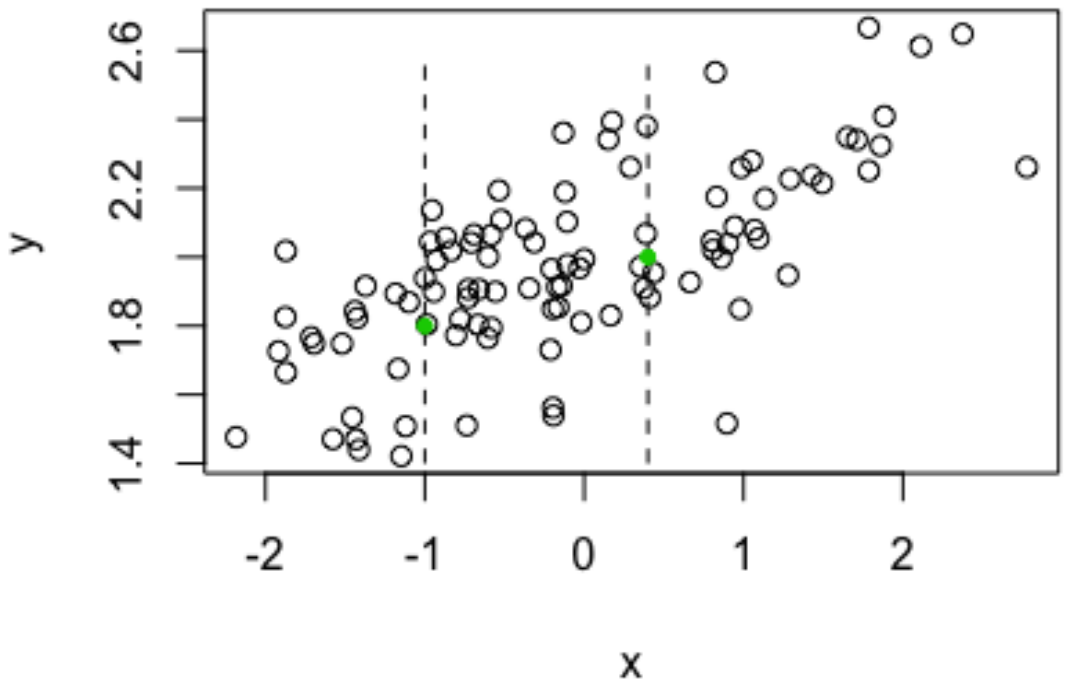
Н.П. Алексеева,

СПбГУ, математико-механический факультет
для 422 группы

2020 г.

Кривая регрессии

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



Экстремальное свойство второго момента

Пусть $\mathrm{m} = \mathbb{E}\xi$, тогда второй момент $\mathbb{E}(\xi - \mathrm{c})^2$ для произвольной константы c имеет минимальное значение при $\mathrm{c} = \mathrm{m}$.

Действительно, обозначим через $\mu_2 = \mathbb{E}(\xi - \mathrm{m})^2$, тогда

$$\mathbb{E}(\xi - \mathrm{c})^2 = \mathbb{E}(\xi - \mathrm{m} + \mathrm{m} - \mathrm{c})^2 =$$
$$= \mu_2 + \underbrace{2(\mathrm{m} - \mathrm{c})\mathbb{E}(\xi - \mathrm{m})}_{=0} + (\mathrm{m} - \mathrm{c})^2 \geq \mu_2$$

Определение и экстремальное свойство регрессии

Пусть ξ и η имеют плотность $f(x,y)$, для любого x частная и условная плотность имеют соответственно вид

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy > 0\,, \quad f(y|x) = \frac{f(x,y)}{f_1(x)}$$

Кривую регрессии величины η определяют как траекторию точки (x,y) при переменном x, где $y = \mathbb{E}(\eta|\xi = x)$.

Среди всех возможных функций $g(\xi)$ минимум выражения

$$\mathbb{E}(\eta - g(\xi))^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (y - g(x))^2 f(x,y)dxdy = \tag{1}$$

$$= \int_{-\infty}^{+\infty} f_1(x)dx \int_{-\infty}^{+\infty} (y - g(x))^2 f(y|x)dy$$

достигается при функции $g(\xi) = \mathbb{E}(\eta|\xi)$.

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

$$y_i = \alpha + \beta x_i + \delta_i \tag{2}$$

где ошибки δ_i такие, что $\mathbb{E}\delta_i = 0, \quad D\delta_i = \sigma^2, \quad \mathbb{E}\delta_i\delta_j = 0$. 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

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

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

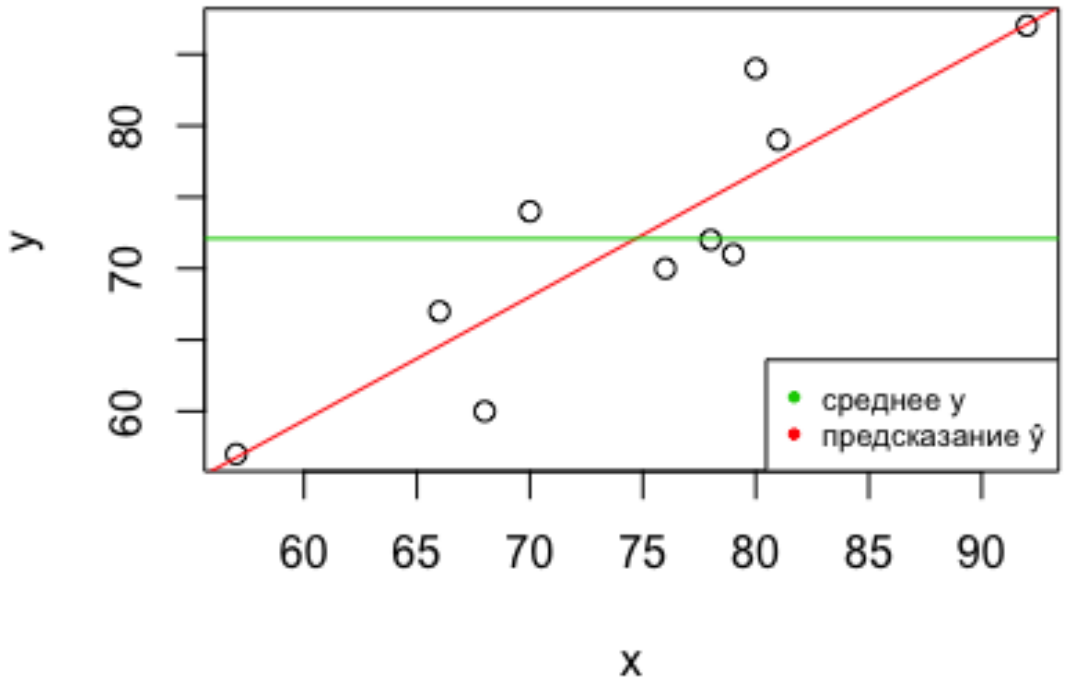


Рис.: $y = 0.8667x + 7.353$.

Некоррелированность остатка с независимой переменной

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}; \qquad \hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

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

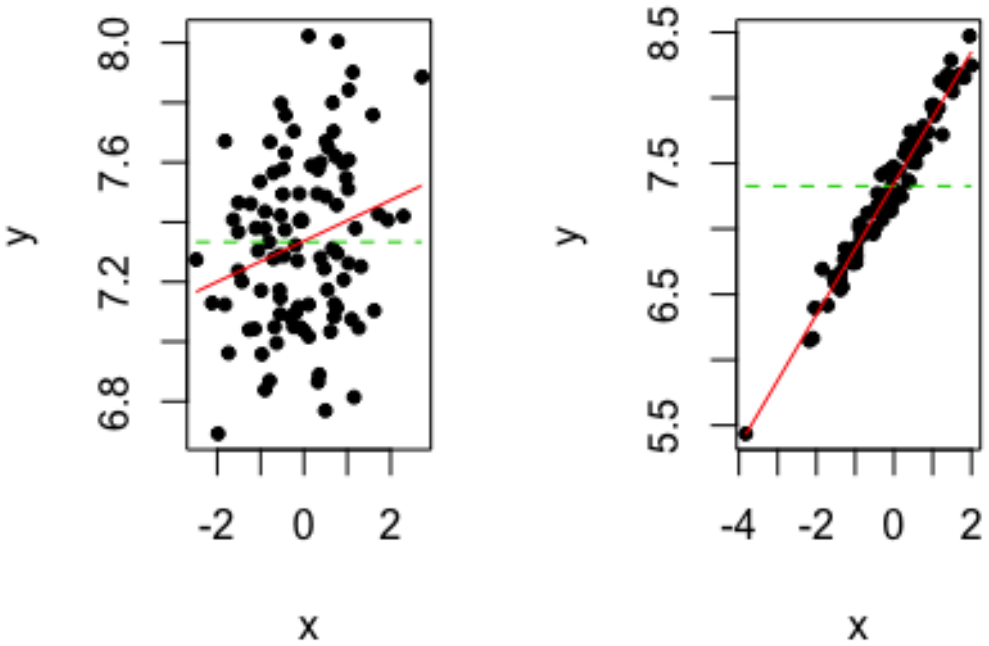
$\frac{1}{n}\sum_{i=1}^{n} \tilde{y}_i = 0$

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

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

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

$$R^2 = 1 - \frac{Q_E}{Q}, \tag{3}$$

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

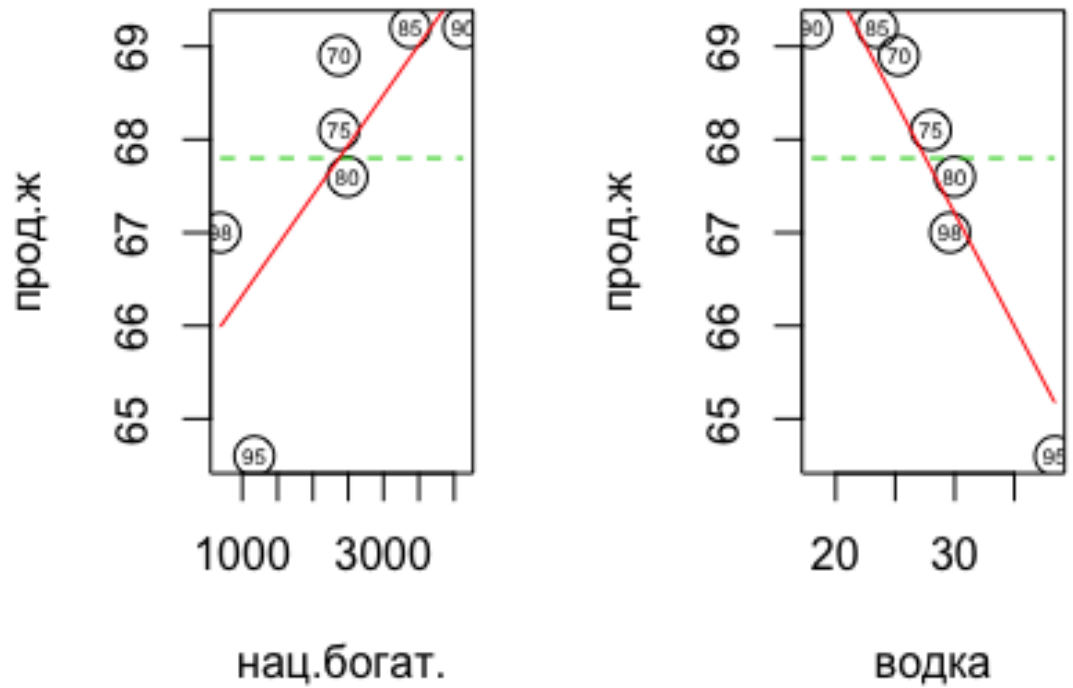


Рис.: $r_1 = 0.776, R_1^2 = 0.6, p = 0.04,$
$r_2 = -0.934, R_2^2 = 0.87, p = 0.002.$

Равенство коэфф.детерминации и квадрата коэфф.корреляции

Выборочный коэффициент корреляции

$$\hat{\varrho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

$$R^2 = 1 - \frac{Q_E}{Q} = \frac{Q_R}{Q} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}.$$

Подставим $\hat{\beta} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$ и сократим на $\sum_{i=1}^{n}(x_i - \bar{x})^2$,

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \frac{\sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} = \hat{\varrho}^2.$$

Вычисление в R

```
> X
```

[1] 2.372 2.372 2.489 3.379 4.130 1.171 0.689

```
> Y
```

[1] 68.9 68.1 67.6 69.2 69.2 64.6 67.0

```
SLM < −summary(lm(Y ∼ X)); coef(SLM);
```

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	65.25	1.02	63.91	0.00
X	1.08	0.39	2.76	0.04

```
> predict(lm(Y ∼  X))
```

67.80 67.80 67.93 68.88 69.69 66.51 65.99

```
> SLM$r.squared
```

[1] 0.6028887

Свойства оценок линейной регрессии

Не умаляя общности, будем считать

$$\sum_{i=1}^{n} x_i = 0. \tag{5}$$

При этом оценки параметров линейной регрессии упрощаются и имеют вид

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{6}$$

Распределение $\hat{\alpha}$ и $\hat{\beta}$ в случае известной дисперсии ошибок σ^2

Пусть ошибки регрессии распределены нормально, то есть $\delta_i \sim \mathcal{N}(0;\sigma)$, $\sum\limits_{i=1}^{n} x_i = 0$. Обозначим через $[x,x]$ скалярное произведение $[x,x] = \sum\limits_{i=1}^{n} x_i^2$. Тогда для оценок коэффициентов регрессии справедливо

$$1)\ \ \hat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma}{\sqrt{n}}\right),$$

$$2)\ \ \hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma}{\sqrt{[x,x]}}\right).$$

Действительно,

$$\hat{\alpha} = \bar{\mathrm{y}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y_i} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\alpha + \beta \mathrm{x_i} + \delta_\mathrm{i}) =$$

$$= \alpha + \frac{\beta}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i} + \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_\mathrm{i}.$$

Так как $\sum\limits_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i} = 0$ по предположению, то получаем

$$\hat{\alpha} = \alpha + \bar{\delta},$$

где $\bar{\delta} = \frac{1}{\mathrm{n}}\sum\limits_{\mathrm{i}=1}^{\mathrm{n}} \delta_\mathrm{i}$. Отсюда $\mathbb{E}\hat{\alpha} = \alpha$, $\mathrm{D}\hat{\alpha} = \sigma^2/\mathrm{n}$, $\hat{\alpha} \sim \mathrm{N}(\alpha, \sigma/\sqrt{\mathrm{n}})$.

Подставим $y_i = \alpha + \beta x_i + \delta_i$ в выражение оценки $\hat{\beta}$.

$$\hat{\beta} = \frac{\sum\limits_i y_i x_i}{[x,x]} = \frac{\sum\limits_i (\alpha + \beta x_i + \delta_i) x_i}{[x,x]} = \frac{\alpha \sum\limits_i x_i + \beta \sum\limits_i x_i^2 + \sum\limits_i \delta_i x_i}{[x,x]}.$$

Из $\sum\limits_{i=1}^{n} x_i = 0$ и $[x,x] = \sum\limits_{i=1}^{n} x_i^2$ получаем

$$\hat{\beta} = \beta + \frac{1}{[x,x]} \sum_i \delta_i x_i, \qquad \mathbb{E}\hat{\beta} = \beta,$$

$$D\hat{\beta} = D\left(\beta + \frac{1}{[x,x]} \sum_i \delta_i x_i\right) = \frac{1}{[x,x]^2} \sum_i x_i^2 D\delta_i = \frac{\sigma^2}{[x,x]}.$$

Заметим, что оценки $\hat{\alpha}$ и $\hat{\beta}$ некоррелированы.

$$\mathbb{E}(\hat{\alpha}-\alpha)(\hat{\beta}-\beta) = \mathbb{E}\left(\frac{1}{n}\sum_i \delta_i \cdot \frac{1}{[x,x]}\sum_j x_j\delta_j\right) =$$

$$= \frac{1}{n[x,x]}\sum_i x_i\mathbb{E}\delta_i^2 = \frac{\sigma^2}{n[x,x]}\sum_i x_i = 0\,.$$

Случай неизвестной дисперсии ошибок σ^2

$Q_E = \sum_i (y_i - \hat{y}_i)^2$ имеет распределение $\sigma^2 \chi^2(n-2)$.

Обозначим через $d_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. Используя то, что $\sum_i x_i = 0$, $\hat{\alpha} = \bar{y}$, $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$, убеждаемся, что

$$\sum_i d_i = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0,$$

$$\sum_i x_i d_i = \sum_i x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0\,.$$

На основании этого

$$\sum_i \delta_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2 =$$

$$= \sum_i ((y_i - \hat{\alpha} - \hat{\beta} x_i) + (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) x_i)^2 =$$

$$= Q_E + n(\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 [x, x].$$

При этом перекрестные суммы оказываются равными нулю:

$$\sum_i d_i (\hat{\alpha} - \alpha) = 0,$$

$$\sum_i d_i x_i (\hat{\beta} - \beta) = 0,$$

$$\sum_i (\hat{\alpha} - \alpha)(\hat{\beta} - \beta) x_i = 0\,.$$

$Q_E \sim \sigma^2\chi^2(n-2)$

$$Q_E = \sum_{i=1}^{n} \delta_i^2 - n(\hat{\alpha}-\alpha)^2 - (\hat{\beta}-\beta)^2[x,x] \sim \sigma^2\chi^2(n-2), \text{ так как}$$

$$y_1 = \sqrt{n}(\hat{\alpha}-\alpha) = \sqrt{n}\bar{\delta} = \frac{1}{\sqrt{n}}(\delta_1 + \ldots + \delta_n),$$

$$y_2 = \sqrt{[x,x]}(\hat{\beta}-\beta) = \frac{1}{\sqrt{[x,x]}} \sum_{i=1}^{n} x_i\delta_i \,.$$

Ортогональная матрица для преобразования $Y = C\delta$ вида

$$C = \begin{bmatrix} \frac{1}{\sqrt{n}} & \ldots & \frac{1}{\sqrt{n}} \\ \frac{x_1}{\sqrt{[x,x]}} & \ldots & \frac{x_n}{\sqrt{[x,x]}} \\ c_{31} & \ldots & c_{3n} \\ \vdots & \ldots & \vdots \\ c_{n1} & \ldots & c_{nn} \end{bmatrix}$$

Используем $\mathbb{E}\delta_i^2 = \sigma^2, \quad \mathbb{E}(\hat\alpha - \alpha)^2 = \dfrac{\sigma^2}{n}, \quad \mathbb{E}(\hat\beta - \beta)^2 = \dfrac{\sigma^2}{[x,x]},$

$$\mathbb{E}(\hat\beta - \beta)(\hat\alpha - \alpha) = 0, \quad \mathbb{E}(\hat\alpha - \alpha)\delta_i = \mathbb{E}\frac{1}{n}\sum_j \delta_j\delta_i = \frac{\sigma^2}{n},$$

$$\mathbb{E}(\hat\beta - \beta)\sum_i \delta_i x_i = \mathbb{E}\frac{1}{[x,x]}\sum_i\sum_j x_j\delta_j\delta_i x_i = \frac{1}{[x,x]}\sum_i \sigma^2 x_i^2 = \sigma^2.$$

$$\mathbb{E}Q_E = \mathbb{E}\sum_i (y_i - \hat\alpha - \hat\beta x_i)^2 = \mathbb{E}\sum_i (\overbrace{\alpha + \beta x_i + \delta_i}^{y_i} - \hat\alpha - \hat\beta x_i)^2 =$$

$$= \mathbb{E}\sum_i (\delta_i - (\hat\alpha - \alpha) - (\hat\beta - \beta)x_i)^2 =$$

$$= \sum_i \mathbb{E}\delta_i^2 + n\mathbb{E}(\hat\alpha - \alpha)^2 + \mathbb{E}(\hat\beta - \beta)^2\sum_i x_i^2 - 2\mathbb{E}(\hat\alpha - \alpha)\sum_i \delta_i -$$

$$-2\mathbb{E}(\hat\beta - \beta)\sum_i \delta_i x_i + 2\mathbb{E}(\hat\alpha - \alpha)(\hat\beta - \beta)\sum_i x_i =$$

$$= n\sigma^2 + \sigma^2 + \sigma^2 - 2\sigma^2 - 2\sigma^2 + 0 = (n-2)\sigma^2.$$

Пусть $\delta_i \sim N(0;\sigma)$, где σ неизвестна, $S^2 = \frac{Q_E}{n-2}$, $S^2_\alpha = S^2/n$, $S^2_\beta = S^2/[x,x]$. Тогда

$$T' = \frac{\hat{\alpha} - \alpha}{S_\alpha} \sim T(n-2), \tag{7}$$

$$T'' = \frac{\hat{\beta} - \beta}{S_\beta} \sim T(n-2), \tag{8}$$

В случае, когда $\sum_{i=1}^{n} x_i \neq 0$, имеем $[x,x] = \sum_{i=1}^{n}(x_i - \bar{x})^2$,

$$S^2_\alpha = \frac{S^2 \sum_{i=1}^{n} x_i^2}{n[x,x]}. \tag{9}$$

Доказательство 1)

Из предыдущего имеем

$$\mathrm{U}_1 = \frac{\hat{\alpha} - \alpha}{\sigma/\sqrt{\mathrm{n}}} \sim \mathrm{N}(0;1),$$
$$\chi^2 = (\mathrm{n}-2)\mathrm{S}^2/\sigma^2 \sim \chi^2(\mathrm{n}-2).$$

Исходя из этого построим случайную величину T', имеющую распределение Стьюдента:

$$\mathrm{T}' = \frac{\mathrm{U}_1}{\sqrt{\frac{1}{\mathrm{n}-2}\chi^2}} = \frac{(\hat{\alpha}-\alpha)\sqrt{\mathrm{n}}}{\sigma \cdot \sqrt{\frac{1}{\mathrm{n}-2}(\mathrm{n}-2)\mathrm{S}^2/\sigma^2}} =$$
$$= \frac{\hat{\alpha}-\alpha}{\mathrm{S}}\sqrt{\mathrm{n}} \sim \mathrm{T}(\mathrm{n}-2).$$

Доказательство 2)

Ранее было получено

$$\mathrm{U}_2 = \frac{\hat{\beta} - \beta}{\sigma / \sqrt{[\mathrm{x}, \mathrm{x}]}} \sim \mathrm{N}(0; 1).$$

Отсюда

$$\mathrm{T}'' = \frac{(\hat{\beta} - \beta)\sqrt{[\mathrm{x}, \mathrm{x}]}}{\sigma \cdot \sqrt{\frac{1}{\mathrm{n}-2}(\mathrm{n}-2)\mathrm{S}^2/\sigma^2}} = \frac{\hat{\beta} - \beta}{\mathrm{S}}\sqrt{[\mathrm{x}, \mathrm{x}]} \sim \mathrm{T}(\mathrm{n}-2).$$

В случае $\bar{\mathrm{x}} \neq 0$ имеем $\hat{\alpha} = \bar{\mathrm{y}} - \hat{\beta}\bar{\mathrm{x}}$, $[\mathrm{x}, \mathrm{x}] = \sum\limits_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2$.

Дисперсия оценки параметра α имеет вид

$$\mathrm{D}\hat{\alpha} = \mathrm{D}(\bar{\mathrm{y}} - \hat{\beta}\bar{\mathrm{x}}) = \mathrm{D}\bar{\mathrm{y}} + \mathrm{D}(\hat{\beta}\bar{\mathrm{x}}) =$$
$$= \frac{\sigma^2}{\mathrm{n}} + \frac{\bar{\mathrm{x}}^2\sigma^2}{[\mathrm{x},\mathrm{x}]} = \sigma^2\frac{\sum \mathrm{x}_\mathrm{i}^2 - \mathrm{n}\bar{\mathrm{x}}^2 + \mathrm{n}\bar{\mathrm{x}}^2}{\mathrm{n}[\mathrm{x},\mathrm{x}]} = \frac{\sigma^2\sum \mathrm{x}_\mathrm{i}^2}{\mathrm{n}[\mathrm{x},\mathrm{x}]}.$$

Таким образом, по выборке дисперсии коэффициентов регрессии могут быть оценены следующим образом:

$$\mathrm{D}\hat{\alpha} = \mathrm{S}_\alpha^2 = \frac{\mathrm{Q}_\mathrm{E}\sum\limits_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_\mathrm{i}^2}{\mathrm{n}(\mathrm{n}-2)\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}, \tag{10}$$

$$\mathrm{D}\hat{\beta} = \mathrm{S}_\beta^2 = \frac{\mathrm{Q}_\mathrm{E}}{(\mathrm{n}-2)\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}. \tag{11}$$

S_α и S_β называются стандартными ошибками коэффициентов регрессии.

Доверительный интервал для линии регрессии

Уравнение линии регрессии $\hat{y} = \hat{\alpha} + \hat{\beta}x$ зависит от случайных величин $\hat{\alpha}$, $\hat{\beta}$, $D\hat{y} = D\hat{\alpha} + x^2 D\hat{\beta}$,

$$D\hat{\alpha} = \frac{\sigma^2 \sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}, \quad D\hat{\beta} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2},$$

$$D\hat{y} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \left(\frac{1}{n} \sum_i x_i^2 + x^2 \right)$$

Для каждого значения x построим статистику Стьюдента

$$t = \frac{\hat{y} - (\alpha + \beta x)}{\sqrt{D\hat{y}} \sqrt{\frac{Q_e}{(n-2)\sigma^2}}} = \frac{\hat{y} - (\alpha + \beta x)}{S}, \quad S^2 = \frac{Q_E \left(\frac{1}{n} \sum_i x_i^2 + x^2 \right)}{(n-2) \sum_i (x_i - \bar{x})^2}$$

Доверительный интервал имеет вид

$$\alpha + \beta x \pm t_{1-\alpha/2}^{(n-2)} S$$

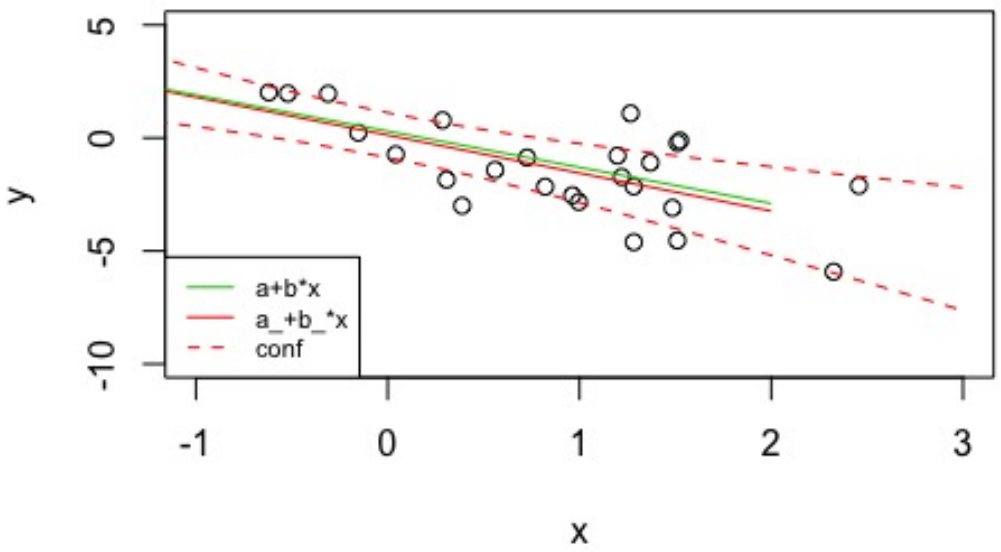


Рис.: Линейная регрессия с доверительным интервалом.

Проверка гипотез в регрессионном анализе

При $H_0 : \beta = 0$ статистика $F = \frac{Q_R}{Q_E/(n-2)} \sim F(1, n-2)$.

Из теоремы Фишера-Кочрена следует, что Q_R и Q_E независимы, $Q_E \sim \chi^2(n-2)$, $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i$

$$Q_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2 = \hat{\beta}^2[x, x].$$

Если $\beta = 0$, то $\hat{\beta} \sim N(0; \sigma/\sqrt{[x, x]})$, поэтому $\xi = \hat{\beta}\sqrt{[x, x]}/\sigma \sim N(0; 1)$, $Q_R/\sigma^2 = \xi^2 \sim \chi^2(1)$.

Статистика $T = \sqrt{F}$ имеет распределение Стьюдента с $n - 2$ степенями свободы. Она выражается через коэффициент детерминации и, соответственно, через коэффициент корреляции.

$$T = \sqrt{F} = \sqrt{\frac{Q_R}{Q_E}}\sqrt{n-2} = \frac{\sqrt{Q_R}}{\sqrt{Q - Q_R}}\sqrt{n-2} = \\ = \frac{\sqrt{Q_R/Q}}{\sqrt{1 - Q_R/Q}}\sqrt{n-2}.$$

Таким образом получаем статистику для проверки значимости коэффициента корреляции

$$T = \frac{\hat{\varrho}}{\sqrt{1 - \hat{\varrho}^2}}\sqrt{n-2} \ \sim \ T(n-2), \qquad (12)$$

которая в случае справедливости нулевой гипотезы $H_0 : \varrho = 0$ имеет распределение Стьюдента с $n - 2$ степенями свободы.