ПРАКТИЧЕСКОЕ ЗАНЯТИЕ 12.

РЕГРЕССИОННЫЙ АНАЛИЗ. ПОСТРОЕНИЕ ЛИНЕЙНОЙ И КВАДРАТИЧНОЙ РЕГРЕССИОННЫХ МОДЕЛЕЙ

1. Задача регрессии для линейной функции.

Рассмотрим случай, когда уравнение регрессии (1.61) является линейной функцией

$$y = \beta_1 + \beta_2 x, \tag{1.69}$$

т.е. базисные функции $\varphi_1(x) = 1$, $\varphi_2(x) = x$. В этом случае система (1.63) имеет вид

$$\begin{cases} \beta_1 n + \beta_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i \\ \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i x_i \end{cases} \tag{1.70}$$

Расчет упростится, если ввести замену $X = \dfrac{x - \overline{x}}{h}$ и рассматривать уравнение

$$y = B_1 + B_2 X = B_1 + B_2 \frac{x - \overline{x}}{h}, \tag{1.71}$$

где $\overline{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$ – среднее арифметическое аргументов *x*, *h* выбирается из условия, чтобы значения *X* были целыми не имеющими общего множителя. Уравнение (1.71) будем называть уравнением с *кодированным переменным*, в отличие от уравнения (1.69) с *реальным переменным*. В этом случае $\sum_{i=1}^{n} X_i = 0$ и система (1.70) будет иметь вид

$$\begin{cases} B_1 n = \sum_{i=1}^{n} Y_i \\ B_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i x_i \end{cases}$$

Откуда имеем формулы для оценок коэффициентов регрессии уравнения с кодированным переменным:

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n}, \qquad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}. \tag{1.72}$$

Для контроля расчетов удобно воспользоваться свойством отклонений $\Delta Y_i = Y_i - \tilde{Y}(x_i)$ экспериментальных результатов Y_i от рассчитанных по оценкам (1.72) значений функции регрессии $\tilde{Y}(x_i) = \tilde{B}_1 + \tilde{B}_2 x_i$:

$$\sum_{i=1}^{n} \Delta Y_i = 0, \tag{1.73}$$

Дисперсия адекватности (1.66) для проверки адекватности линейной регрессионной модели вычисляется по формуле

$$S_{\text{ад}}^2 = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{\text{ад}}}; \quad k_{\text{ад}} = n - 2. \tag{1.74}$$

Границы доверительных интервалов для параметров линейной функции регрессии с кодированным переменным (1.71) имеют вид

$$\tilde{B}_1 \pm \varepsilon_1; \qquad \varepsilon_1 = t_{1-\frac{\alpha}{2}}(k) \frac{S}{\sqrt{n}}; \qquad\qquad \tilde{B}_2 \pm \varepsilon_2; \qquad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(k) \frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2}}. \tag{1.75}$$

Оценки коэффициентов регрессии линейной функции (1.69) с реальным переменным при этом могут быть найдены по формулам:

$$\tilde{\beta}_1 = \tilde{B}_1 - \tilde{B}_2 \frac{\overline{x}}{h}; \qquad\qquad \tilde{\beta}_2 = \frac{\tilde{B}_2}{h}. \tag{1.76}$$

Границы доверительных интервалов для коэффициентов линейной функции с реальным переменным (1.69) имеют вид

$$\tilde{\beta}_1 \pm \hat{\varepsilon}_1; \qquad \hat{\varepsilon}_1 = t_{1-\frac{\alpha}{2}}(k) S \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{h^2 \sum_{i=1}^{n} X_i^2}};$$

$$\tilde{\beta}_2 \pm \hat{\varepsilon}_2; \qquad \hat{\varepsilon}_2 = t_{1-\frac{\alpha}{2}}(k) \frac{S}{h\sqrt{\sum_{i=1}^{n} X_i^2}}. \tag{1.77}$$

2. Задача регрессии для квадратичной функции.

Рассмотрим случай, когда уравнение регрессии (1.61) является квадратичной функцией

$$y = \beta_1 + \beta_2 x + \beta_3 x^2, \tag{1.78}$$

т.е. базисные функции $\varphi_1(x) = 1$, $\varphi_2(x) = x$, $\varphi_3(x) = x^2$. В этом случае система (1.63) имеет вид

$$\begin{cases} \beta_1 n + \beta_2 \sum_{i=1}^{n} x_i + \beta_3 \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} Y_i \\ \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^{\,2} + \beta_3 \sum_{i=1}^{n} x_i^{\,3} = \sum_{i=1}^{n} Y_i x_i \\ \beta_1 \sum_{i=1}^{n} x_i^{\,2} + \beta_2 \sum_{i=1}^{n} x_i^{\,3} + \beta_3 \sum_{i=1}^{n} x_i^{\,4} = \sum_{i=1}^{n} Y_i x_i^{\,2} \end{cases} \tag{1.79}$$

Как и в предыдущем параграфе, сделаем замену $X = \dfrac{x - \overline{x}}{h}$, где $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ - среднее арифметическое аргументов *x*, *h* выбираем из условия, чтобы значения *X* были целыми, не имеющими общего множителя. Т.е. преобразуем уравнение (1.78) к виду

$$y = B_1 + B_2 X + B_3 X^2 = B_1 + B_2 \frac{x - \overline{x}}{h} + B_3 \left(\frac{x - \overline{x}}{h} \right)^2. \tag{1.80}$$

В этом случае $\sum_{i=1}^{n} X_i = 0$. Кроме того, введем условие $\sum_{i=1}^{n} X_i^{\,3} = 0$. Оно будет выполняться, например, в том случае, когда переменная *x* в исходных данных меняется с постоянным шагом. Тогда система (1.79) существенно упрощается:

$$\begin{cases} B_1 n + B_3 \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} Y_i \\ B_2 \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} Y_i x_i \\ B_1 \sum_{i=1}^{n} x_i^{\,2} + B_3 \sum_{i=1}^{n} x_i^{\,4} = \sum_{i=1}^{n} Y_i x_i^{\,2} \end{cases} \tag{1.81}$$

Решая эту систему, получаем формулы для оценок коэффициентов регрессии уравнения с кодированным переменным (1.80):

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} X_i^{\,4} \cdot \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i^{\,2} \cdot \sum_{i=1}^{n} Y_i X_i^2}{n \sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2} \right)^2}; \quad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^{\,2}}; \quad \tilde{B}_3 = \frac{n \sum_{i=1}^{n} Y_i X_i^2 \cdot - \sum_{i=1}^{n} X_i^{\,2} \cdot \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2} \right)^2}. \tag{1.82}$$

Формула (1.73): $\sum_{i=1}^{n} \Delta Y_i = 0$ для квадратичной модели регрессии также имеет место, и ее удобно использовать для контроля расчетов. Дисперсия адекватности (1.66) для проверки адекватности квадратичной функции регрессии вычисляется по формуле

$$S_{\text{ад}}^2 = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{\text{ад}}}; \quad k_{\text{ад}} = n - 3. \tag{1.83}$$

Границы доверительных интервалов для параметров квадратичной функции регрессии с кодированным переменным (1.80) имеют вид

$$\tilde{B}_1 \pm \varepsilon_1; \qquad \varepsilon_1 = t_{1-\frac{\alpha}{2}}(k)\, S \sqrt{\frac{\sum\limits_{i=1}^{n} X_i^{\,4}}{n\sum\limits_{i=1}^{n} X_i^{\,4} - \left(\sum\limits_{i=1}^{n} X_i^{\,2}\right)^2}};$$

$$\tilde{B}_2 \pm \varepsilon_2; \qquad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{\sqrt{\sum\limits_{i=1}^{n} X_i^{\,2}}}. \tag{1.84}$$

$$\tilde{B}_3 \pm \varepsilon_3; \qquad \varepsilon_3 = t_{1-\frac{\alpha}{2}}(k)\, S \sqrt{\frac{n}{n\sum\limits_{i=1}^{n} X_i^{\,4} - \left(\sum\limits_{i=1}^{n} X_i^{\,2}\right)^2}}.$$

Оценки коэффициентов регрессии квадратичной функции с реальным переменным (1.78) при этом будут

$$\tilde{\beta}_1 = \tilde{B}_1 - \tilde{B}_2\frac{\overline{x}}{h} + \tilde{B}_3\left(\frac{\overline{x}}{h}\right)^2; \qquad \tilde{\beta}_2 = \frac{\tilde{B}_2}{h} - \frac{2\tilde{B}_3\overline{x}}{h^2}; \qquad \tilde{\beta}_3 = \frac{\tilde{B}_3}{h^2}. \tag{1.85}$$

Границы доверительных интервалов для параметров квадратичной функции с реальным переменным (1.78) имеют вид

$$\tilde{\beta}_1 \pm \hat{\varepsilon}_1; \qquad \hat{\varepsilon}_1 = t_{1-\frac{\alpha}{2}}(k)\, S \sqrt{\frac{\sum\limits_{i=1}^{n} X_i^{\,4} + \frac{\overline{x}^4}{h^4}n}{n\sum\limits_{i=1}^{n} X_i^{\,4} - \left(\sum\limits_{i=1}^{n} X_i^{\,2}\right)^2} + \frac{\overline{x}^2}{h^2\sum\limits_{i=1}^{n} X_i^{\,2}}};$$

$$\tilde{\beta}_2 \pm \hat{\varepsilon}_2; \qquad \hat{\varepsilon}_2 = t_{1-\frac{\alpha}{2}}(k)\,\frac{\mathrm{S}}{h}\sqrt{\frac{1}{\sum\limits_{i=1}^{n} X_i^{\,2}} + \frac{4n\overline{x}^2}{h^2\left(n\sum\limits_{i=1}^{n} X_i^{\,4} - \left(\sum\limits_{i=1}^{n} X_i^{\,2}\right)^2\right)}}; \tag{1.86}$$

$$\tilde{\beta}_3 \pm \hat{\varepsilon}_3; \qquad \hat{\varepsilon}_3 = t_{1-\frac{\alpha}{2}}(k)\,\frac{\mathrm{S}}{h^2}\sqrt{\frac{n}{n\sum\limits_{i=1}^{n} X_i^{\,4} - \left(\sum\limits_{i=1}^{n} X_i^{\,2}\right)^2}}$$

Задача 1.25 Результаты экспериментов представлены в первых двух столбцах табл. 1.11. Экспериментальные значения *Y* являются независимыми и равноточными. Построить линейную и квадратичную регрессионные модели.

Таблица 1.11

Условие задачи 1.25 и результаты расчета

	x	Y	$X=\frac{x-0,6}{0,2}$	$X{\cdot}Y$	X^2	$Y_{лин}$	$\Delta Y_{лин}$	$\Delta Y_{лин}^2$
	0,2	4,5	–2	–9,0	4	5,3	–0,8	0,64
	0,4	7,0	–1	–7,0	1	6,25	0,75	0,56
	0,6	8,0	0	0,0	0	7,2	0,8	0,64
	0,8	7,5	1	7,5	1	8,15	–0,65	0,42
	1,0	9,0	2	18,0	4	9,1	–0,1	0,04
Σ	3,0	36,0	0	9,5	10		0	2,30

Решение:

Сначала найдем решение задачи регрессии в кодированных значениях переменной x (1.71). Введем новую переменную по формуле $X=\dfrac{x-\overline{x}}{h}$, где $\overline{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i=\dfrac{1}{5}\cdot 3=0,6.$ Если значения величины x – 0,6 поделить на число h = 0,2, то получатся целые значения, не имеющие общего множителя. Поэтому $X=\dfrac{x-0,6}{0,2}$.

По формулам (1.72) находим оценки коэффициентов линейной регрессии

$$\tilde{B}_1=\frac{\sum_{i=1}^{n}Y_i}{n}=\frac{36}{5}=7,2; \qquad \tilde{B}_2=\frac{\sum_{i=1}^{n}Y_iX_i}{\sum_{i=1}^{n}X_i^{\,2}}=\frac{9,5}{10}=0,95.$$

Получили линейную модель регрессии

$$\widehat{Y}_{лин}=7,2+0,95X.$$

По полученной формуле вычисляем значения линейной функции регрессии $\widehat{Y}_{лин}$ при всех значениях аргумента X, а затем рассчитываем $\Delta Y_i=\widehat{Y}_i-\widehat{Y}_{i\,лин}$ отклонения экспериментальных значений Y_i от значений $\widehat{Y}_{i\,лин}$, полученных по функции регрессии. Контроль, согласно формуле $\sum_{i=1}^{n}\Delta Y_i=0$, выполнен. Все расчеты приведены в таблице 1.11.

Уравнение линейной регрессии Y от реального переменного x найдем, сделав преобразование:

$$Y_{лин}=\beta_1+\beta_2x=7,2+0,95\frac{x-0,6}{0,2}=4,35+4,75x.$$

Построим квадратичную модель регрессии (1.80)

$$y = B_1 + B_2 X + B_3 X^2 = B_1 + B_2 \frac{x - 0,6}{0,2} + B_3 \left(\frac{x - 0,6}{0,2} \right)^2 .$$

Заметим, что в нашем случае $\sum_{i=1}^{n} X_i^3 = 0$, поэтому для оценок коэффициентов регрессии можно воспользоваться формулами (1.82). Все расчеты приведены в таблице 1.12.

Таблица 1.12.

Результаты расчета квадратичной модели регрессии (к задаче 1.25)

	X^3	YX^2	X^4	$Y_{кв}$	$\Delta Y_{кв}$	$\Delta Y_{кв}^2$
	–8	18	16	4,8	–0,3	0,09
	–1	7	1	6,5	0,5	0,25
	0	0	0	7,7	0,3	0,09
	1	7,5	1	8,4	–0,9	0,81
	8	36	16	8,6	0,4	0,16
Σ	0	68,5	34		0	1,4

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} X_i^4 \cdot \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i^2 \cdot \sum_{i=1}^{n} Y_i X_i^2}{n \sum_{i=1}^{n} X_i^4 - \left(\sum_{i=1}^{n} X_i^2 \right)^2} = \frac{34 \cdot 36 - 10 \cdot 68,5}{5 \cdot 34 - 100} = 7,7 ;$$

$$\tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} = \frac{9,5}{10} = 0,95;$$

$$\tilde{B}_3 = \frac{n \sum_{i=1}^{n} Y_i X_i^2 \cdot - \sum_{i=1}^{n} X_i^2 \cdot \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^4 - \left(\sum_{i=1}^{n} X_i^2 \right)^2} = \frac{5 \cdot 68,5 - 10 \cdot 36}{5 \cdot 34 - 100} = -0,25 .$$

Получили квадратичную модель регрессии

$$\widehat{Y}_{кв} = 7,7 + 0,95X - 0,25X^2.$$

Условие $\sum_{i=1}^{n} \Delta Y_i = 0$, выполняется.

Уравнение квадратичной регрессии Y от реального переменного x найдем, сделав преобразование:

$$Y_{кв} = \beta_1 + \beta_2 x + \beta_3 x^2 = 7,7 + 0,95\frac{x - 0,6}{0,2} - 0,25\left(\frac{x - 0,6}{0,2}\right)^2 = 2,6 + 12,25x - 6,25x^2.$$

Пригодность каждой из полученных регрессионных моделей будет оценена на следующем занятии.