4. Обработка данных методами регрессионного анализа

4.1. ТЕОРЕТИЧЕСКОЕ ВВЕДЕНИЕ.

4.1.1. Оценка коэффициентов регрессии.

Важной задачей математической статистики является получение функциональной зависимости одной величины (y) от другой (x) по результатам эксперимента. Будем считать, что функциональная зависимость между величинами, называемая в дальнейшем *моделью*, известна из предварительных сведений с точностью до параметров $\beta_1,\beta_2,...,\beta_m$ и имеет вид

$$y = f(x,\beta_1,\beta_2,...,\beta_m). \tag{4.1}$$

Для отыскания неизвестных параметров проведено n наблюдений (x_i, Y_i), $i = 1, 2, ..., n$. Но так как результаты наблюдений не свободны от погрешностей измерений, которые мы будем рассматривать как случайные ошибки, то по ним нельзя точно найти искомые параметры. Поэтому приходится ставить задачу об отыскании не значений параметров, а их оценок по результатам эксперимента.

Будем предполагать, что значения аргументов x_i известны точно, а значения функции Y_i – взаимно независимые случайные величины, включающие случайные ошибки Z_i, т.е. $Y_i = f(x_i,\beta_1,\beta_2,...,\beta_m) + Z_i$, где

$$M(Z_i) = 0; \quad D(Z_i) = D(Y_i) = \sigma^2.$$

Здесь мы предполагаем, что измерения *равноточны*. Для оценок параметров $\beta_1,\beta_2,...,\beta_m$ используется *метод наименьших квадратов*. В качестве оценок этих параметров принимаются значения $\tilde{\beta}_1,\tilde{\beta}_2,...,\tilde{\beta}_m$, при которых имеет минимум функция (МНК-оценки)

$$Q(\beta_1,\beta_2,...,\beta_m) = \sum_{i=1}^{n} (Y_i - f(x_i,\beta_1,\beta_2,...,\beta_m))^2. \tag{4.2}$$

Уравнение (4.1) называется *уравнением регрессии*, а отыскание оценок параметров и исследование получаемых моделей – *регрессионным анализом*.

Будем рассматривать уравнения регрессии, линейные относительно оцениваемых параметров $\beta_1,\beta_2,...,\beta_m$:

$$y = f(x,\beta_1,\beta_2,...,\beta_m) = \beta_1\varphi_1(x) + \beta_2\varphi_2(x) + ... + \beta_m\varphi_m(x). \tag{4.3}$$

Функции $\varphi_1(x), \varphi_2(x), ..., \varphi_m(x)$ называются *базисными функциями*, их рассматривают на множестве точек $\{x_1, x_2, ..., x_n\}$, где n – число экспериментов. Функция Q (4.2) в этом случае запишется в виде:

$$Q(\beta_1, \beta_2, ..., \beta_m) = \sum_{i=1}^{n} (Y_i - \beta_1\varphi_1(x_i) - \beta_2\varphi_2(x_i) - ... - \beta_m\varphi_m(x_i))^2. \quad (4.4)$$

Для нахождения минимума найдем частные производные функции $Q(\beta_1, \beta_2, ..., \beta_m)$ по переменным $\beta_1, \beta_2, ..., \beta_m$ и приравняем их к нулю (необходимые условия минимума функции). Получим систему уравнений:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} (Y_i - \beta_1\varphi_1(x_i) - \beta_2\varphi_2(x_i) - ... - \beta_m\varphi_m(x_i))\varphi_1(x_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_2} = -2\sum_{i=1}^{n} (Y_i - \beta_1\varphi_1(x_i) - \beta_2\varphi_2(x_i) - ... - \beta_m\varphi_m(x_i))\varphi_2(x_i) = 0 \\ ..... \\ \dfrac{\partial Q}{\partial \beta_m} = -2\sum_{i=1}^{n} (Y_i - \beta_1\varphi_1(x_i) - \beta_2\varphi_2(x_i) - ... - \beta_m\varphi_m(x_i))\varphi_m(x_i) = 0 \end{cases}$$

которую после преобразований можно записать в виде:

$$\begin{cases} \beta_1\sum_{i=1}^{n} \varphi_1^2(x_i) + \beta_2\sum_{i=1}^{n} \varphi_1(x_i)\varphi_2(x_i) + ... + \beta_m\sum_{i=1}^{n} \varphi_1(x_i)\varphi_m(x_i) = \sum_{i=1}^{n} Y_i\varphi_1(x_i) \\ \beta_1\sum_{i=1}^{n} \varphi_1(x_i)\varphi_2(x_i) + \beta_2\sum_{i=1}^{n} \varphi_2^2(x_i) + ... + \beta_m\sum_{i=1}^{n} \varphi_2(x_i)\varphi_m(x_i) = \sum_{i=1}^{n} Y_i\varphi_2(x_i) \\ ..... \\ \beta_1\sum_{i=1}^{n} \varphi_1(x_i)\varphi_m(x_i) + \beta_2\sum_{i=1}^{n} \varphi_2(x_i)\varphi_m(x_i) + ... + \beta_m\sum_{i=1}^{n} \varphi_m^2(x_i) = \sum_{i=1}^{n} Y_i\varphi_m(x_i) \end{cases} \quad (4.5)$$

Следовательно, оценки параметров $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_m$ являются решениями линейной алгебраической системы m уравнений (4.5).

Введем обозначения: $\sum_{i=1}^{n} \varphi_j(x_i)\varphi_k(x_i) = (\varphi_j, \varphi_k)$; $\sum_{i=1}^{n} Y_i\varphi_k(x_i) = (Y, \varphi_k)$, тогда система (4.5) запишется в виде:

$$\begin{cases} \beta_1(\varphi_1, \varphi_1) + \beta_2(\varphi_1, \varphi_2) + ... + \beta_m(\varphi_1, \varphi_m) = (Y, \varphi_1) \\ \beta_1(\varphi_1, \varphi_2) + \beta_2(\varphi_2, \varphi_2) + ... + \beta_m(\varphi_2, \varphi_m) = (Y, \varphi_2) \\ ..... \\ \beta_1(\varphi_1, \varphi_m) + \beta_2(\varphi_2, \varphi_m) + ... + \beta_m(\varphi_m, \varphi_m) = (Y, \varphi_m) \end{cases}$$

С использованием следующих матричных обозначений:

$$A=\begin{pmatrix}(\varphi_1,\varphi_1) & (\varphi_1,\varphi_2) & \dots & (\varphi_1,\varphi_m)\\ (\varphi_1,\varphi_2) & (\varphi_2,\varphi_2) & \dots & (\varphi_2,\varphi_m)\\ \dots & \dots & \dots & \dots\\ (\varphi_1,\varphi_m) & (\varphi_2,\varphi_m) & \dots & (\varphi_m,\varphi_m)\end{pmatrix}$$ - матрица коэффициентов при неизвестных,

$$Y=\begin{pmatrix}(Y,\varphi_1)\\ (Y,\varphi_2)\\ \dots\\ (Y,\varphi_m)\end{pmatrix}$$ - вектор правых частей, $$\beta=\begin{pmatrix}\beta_1\\ \beta_2\\ \dots\\ \beta_m\end{pmatrix}$$ - вектор параметров,

Система (4.5) принимает вид

$$A\beta=Y\,. \tag{4.6}$$

При условии, что A – невырожденная матрица, решение системы (4.6) можно записать в виде

$$\tilde{\beta}=A^{-1}Y\,, \tag{4.7}$$

где $\tilde{\beta}=\begin{pmatrix}\tilde{\beta}_1\\ \tilde{\beta}_2\\ \dots\\ \tilde{\beta}_m\end{pmatrix}$ – вектор МНК-оценок параметров регрессионной модели (4.3).

Оценки параметров линейной регрессии, получаемые по методу наименьших квадратов, имеют следующие свойства:

1. Они являются линейными функциями результатов наблюдений $Y_i, i=1,2,...,n$, и несмещенными оценками параметров, т.е. $M(\tilde{\beta}_j)=\beta_j, \quad j=1,2,...,m.$
2. Они имеют минимальные дисперсии в классе несмещенных оценок, являющихся линейными функциями результатов наблюдений.

4.1.2. Построение доверительных интервалов для коэффициентов регрессии.

Оценки параметров регрессии $\tilde{\beta}_j$, определяемые формулами (4.7), являются точечными оценками истинных значений параметров β_j. Если результаты экспериментов независимы и подчиняются нормальному закону распределения с дисперсией σ^2, то доверительный интервал с доверительной вероятностью $P=1-\alpha$ для каждого параметра $\tilde{\beta}_j$ можно определить неравенством

$$\left|B_j-\widetilde{B}_j\right|<\varepsilon_j \quad \text{или} \quad B_j=\widetilde{B}_j\pm\varepsilon_j\,, \tag{4.8}$$

где

$$\varepsilon_j = t_{1-\alpha/2}(k) S\sqrt{a_{jj}}\,; \qquad (j = 1, 2, ..., m.), \tag{4.9}$$

здесь S^2 – несмещенная оценка дисперсии σ^2 с числом степеней свободы k; $t_{1-\alpha/2}(k)$ – квантиль распределения Стьюдента; a_{jj} – диагональный элемент матрицы A^{-1}, $\alpha = 1 - P$, P – доверительная вероятность.

4.1.3. Проверка гипотезы об адекватности регрессионной модели.

Регрессионная модель называется *адекватной*, если предсказанные по ней значения переменной Y согласуются с результатами эксперимента. Если модель адекватна, то отклонения результатов эксперимента от полученной функции регрессии $\Delta Y_i = Y_i - \widetilde{Y}(x_i)$ являются реализациями случайных ошибок эксперимента Z_i ,которые, в силу предположений (4.1), должны быть независимыми нормально распределенными случайными величинами с нулевыми средними и одинаковыми дисперсиями σ^2. Проверка выполнения этих предположений осуществляются статистическими методами и лежит в основе оценки адекватности модели регрессии.

Для проверки адекватности регрессионной модели вычисляют остаточную дисперсию (так называемую дисперсию адекватности) по формуле

$$S^2_{\text{ад}} = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{\text{ад}}}; \quad k_{\text{ад}} = n - m, \tag{4.10}$$

где ΔY_i – отклонения средних Y_i от проверяемой модели регрессии; $k_{\text{ад}}$ – число степеней свободы дисперсии адекватности; n – число точек, в которых проводился эксперимент; m – число оцениваемых параметров β_j в проверяемой модели.

Если истинная функция регрессии имеет тот же вид, что и рассматриваемая модель (например, так же, как и модель, представляет собой квадратичную функцию), то дисперсия адекватности служит несмещенной оценкой истинной дисперсии эксперимента и ее можно сравнивать с другими подобными оценками. В частности, может быть проведена независимая серия измерений для получения оценки дисперсии эксперимента $S^2_{\text{экс}}$. В этом случае $S^2_{\text{экс}}$ оценивает дисперсию эксперимента $D_{\text{экс}}$, $S^2_{\text{ад}}$ характеризует степень отклонения экспериментальных точек от регрессионной модели, т.е. оценивает некую дисперсию адекватности $D_{\text{ад}}$. Проверка адекватности модели заключается в проверке гипотезы H_0: $D_{\text{ад}} = D_{\text{экс}}$ при альтернативной гипотезе H_1: $D_{\text{ад}} > D_{\text{экс}}$ (если модель неадекватна, отклонения экспериментальных точек от модели будут больше погрешностей эксперимента). Таким образом, задача сводится к проверке гипотезы о равенстве дисперсий, которая решается с помощью критерия Фишера. Вычисляем отношение

$$F = S^2_{\text{ад}} / S^2_{\text{экс}}. \tag{4.11}$$

Если при заданном уровне значимости α отношение F окажется меньше квантили $F_{1-\alpha}(k_1, k_2)$, где $k_1 = k_{\text{ад}}$, $k_2 = k_{\text{экс}}$, то рассматриваемая модель не противоречит результатам эксперимента и принимается; в противоположном случае модель отвергается с уровнем значимости α, как противоречащая результатам эксперимента.

В построенной регрессионной модели (4.3) некоторые коэффициенты могут быть незначимы, т.е. может выполняться гипотеза H_0: $\beta_j = 0$. Для проверки этой гипотезы можно найти доверительный интервал для коэффициента β_j с уровнем значимости α. Если этот интервал «накрывает» значение $\beta_j = 0$, гипотеза H_0 принимается и коэффициент β_j признается незначимым, в противоположном случае коэффициент β_j значим.

4.1.4. Задача регрессии для линейной функции.

Рассмотрим случай, когда уравнение регрессии (4.3) является линейной функцией

$$y = \beta_1 + \beta_2 x, \tag{4.12}$$

т.е. базисные функции $\varphi_1(x) = 1$, $\varphi_2(x) = x$. В этом случае система (4.5) имеет вид

$$\begin{cases} \beta_1 n + \beta_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i \\ \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i x_i \end{cases} \tag{4.13}$$

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

$$y = B_1 + B_2 X = B_1 + B_2 \frac{x - \overline{x}}{h}, \tag{4.14}$$

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 (4.14) будем называть уравнением с *кодированным переменным*, в отличие от уравнения (4.12) с *реальным переменным*. В этом случае $\sum_{i=1}^{n} X_i = 0$ и система (4.13) 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 (4.14) с кодированным переменным:

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n}, \qquad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^{\,2}}. \tag{4.15}$$

Для контроля расчетов удобно воспользоваться свойством отклонений $\Delta Y_i = Y_i - \tilde{Y}(x_i)$ экспериментальных результатов Y_i от рассчитанных по оценкам (4.15) значений функции регрессии $\tilde{Y}(x_i) = \tilde{B}_1 + \tilde{B}_2 x_i$:

$$\sum_{i=1}^{n} \Delta Y_i = 0, \tag{4.16}$$

Дисперсия адекватности (4.10) для проверки адекватности линейной регрессионной модели вычисляется по формуле

$$S_{\text{ад}}^2 = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{\text{ад}}}; \quad k_{\text{ад}} = n - 2. \tag{4.17}$$

Границы доверительных интервалов для параметров линейной функции регрессии с кодированным переменным (4.14) имеют вид

$$\tilde{B}_1 \pm \varepsilon_1; \qquad \varepsilon_1 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{\sqrt{n}}; \qquad \tilde{B}_2 \pm \varepsilon_2; \qquad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{\sqrt{\sum_{i=1}^{n} X_i^{\,2}}}. \tag{4.18}$$

Оценки коэффициентов регрессии линейной функции (4.12) с реальным переменным при этом могут быть найдены по формулам:

$$\tilde{\beta}_1 = \tilde{B}_1 - \tilde{B}_2 \frac{\overline{x}}{h}; \qquad \tilde{\beta}_2 = \frac{\tilde{B}_2}{h}. \tag{4.19}$$

Границы доверительных интервалов для коэффициентов линейной функции с реальным переменным (4.12) имеют вид

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

$$\tilde{\beta}_2 \pm \hat{\varepsilon}_2; \qquad \hat{\varepsilon}_2 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{h\sqrt{\sum_{i=1}^{n} X_i^{\,2}}}. \tag{4.20}$$

4.1.5. Задача регрессии для квадратичной функции.

Рассмотрим случай, когда уравнение регрессии (4.3) является квадратичной функцией

$$y = \beta_1 + \beta_2 x + \beta_3 x^2, \tag{4.21}$$

т.е. базисные функции $\varphi_1(x) = 1$, $\varphi_2(x) = x$, $\varphi_3(x) = x^2$. В этом случае система (4.5) имеет вид

$$\begin{cases} \beta_1 n + \beta_2 \sum_{i=1}^{n} x_i + \beta_3 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i \\ \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 + \beta_3 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} Y_i x_i \\ \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 + \beta_3 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} Y_i x_i^2 \end{cases} \quad (4.22)$$

Как и в предыдущем параграфе, сделаем замену $X = \frac{x - \overline{x}}{h}$, где $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ - среднее арифметическое аргументов *x*, *h* выбираем из условия, чтобы значения *X* были целыми не имеющими общего множителя. Т.е. преобразуем уравнение (4.21) к виду

$$y = B_1 + B_2 X + B_3 X^2 = B_1 + B_2 \frac{x - \overline{x}}{h} + B_3 \left(\frac{x - \overline{x}}{h} \right)^2 . \quad (4.23)$$

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 (4.22) существенно упрощается:

$$\begin{cases} B_1 n + B_3 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i \\ B_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i x_i \\ B_1 \sum_{i=1}^{n} x_i^2 + B_3 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} Y_i x_i^2 \end{cases} \quad (4.24)$$

Решая эту систему, получаем формулы для оценок коэффициентов регрессии уравнения с кодированным переменным (4.23):

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} X_i^4 \cdot \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i^2 \cdot \sum_{i=1}^{n} Y_i X_i^2}{n \sum_{i=1}^{n} X_i^4 - \left(\sum_{i=1}^{n} X_i^2 \right)^2}; \qquad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}; \qquad \tilde{B}_3 = \frac{n \sum_{i=1}^{n} Y_i X_i^2 \cdot - \sum_{i=1}^{n} X_i^2 \cdot \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^4 - \left(\sum_{i=1}^{n} X_i^2 \right)^2}. \quad (4.25)$$

Формула (4.16): $\sum_{i=1}^{n} \Delta Y_i = 0$ для квадратичной модели регрессии также имеет место, и ее удобно использовать для контроля расчетов. Дисперсия адекватности (4.10) для проверки адекватности квадратичной функции регрессии вычисляется по формуле

$$S_{\text{ад}}^2 = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{\text{ад}}}; \quad k_{\text{ад}} = n - 3. \quad (4.26)$$

Границы доверительных интервалов для параметров квадратичной функции регрессии с кодированным переменным (4.23) имеют вид

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

$$\tilde{B}_2 \pm \varepsilon_2; \qquad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(k) \frac{S}{\sqrt{\sum_{i=1}^{n} X_i^{\,2}}}. \tag{4.27}$$

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

Оценки коэффициентов регрессии квадратичной функции с реальным переменным (4.20) при этом будут

$$\tilde{\beta}_1 = \tilde{B}_1 - \tilde{B}_2 \frac{\overline{x}}{h} + \tilde{B}_3 \left(\frac{\overline{x}}{h}\right)^2; \qquad \tilde{\beta}_2 = \frac{\tilde{B}_2}{h} - \frac{2\tilde{B}_3 \overline{x}}{h^2}; \qquad \tilde{\beta}_3 = \frac{\tilde{B}_3}{h^2}. \tag{4.28}$$

Границы доверительных интервалов для параметров квадратичной функции с реальным переменным (4.20) имеют вид

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

$$\tilde{\beta}_2 \pm \hat{\varepsilon}_2; \qquad \hat{\varepsilon}_2 = t_{1-\frac{\alpha}{2}}(k)\, \frac{\mathrm{S}}{h} \sqrt{\frac{1}{\sum_{i=1}^{n} X_i^{\,2}} + \frac{4n\overline{x}^2}{h^2\left(n\sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2}\right)^2\right)}}; \tag{4.29}$$

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

4.2. Содержание типового расчета.

Типовой расчет состоит из двух задач. В каждой задаче приведены результаты независимых равноточных экспериментов по изучению зависимости одной величины (y) от другой (x). В каждой серии проведено n наблюдений (x_i, Y_i), $i = 1, 2, \ldots, n$. Будем предполагать, что значения аргументов x_i известны точно, а значения функции Y_i – взаимно независимые случайные

величины $Y_i = f(x_i) + Z_i$, где погрешности эксперимента Z_i - независимые равноточные случайные величины, имеющие нормальное распределение

$$M(Z_i) = 0; \quad D(Y_i) = \sigma^2.$$

По отдельной серии измерений найдена несмещенная оценка дисперсии S^2.

Задача 1. По приведенным исходным данным требуется:

– найти оценки для параметров линейной регрессии *Y* на *x*, проверить адекватность построенной модели; приняв уровень значимости α = 0,05;

– найти границы доверительных интервалов для параметров линейной модели с доверительной вероятностью *P* = 0,95;

– построить график полученной модели регрессии и график отклонений функции регрессии от экспериментальных данных.

Задача 2. По приведенным исходным данным требуется:

– найти оценки для параметров линейной регрессии *Y* на *x*; проверить адекватность построенной модели; приняв уровень значимости α = 0,05;

– если линейная модель адекватна, найти границы доверительных интервалов для параметров полученной линейной модели с доверительной вероятностью *P* = 0,95;

– если линейная модель неадекватна, найти оценки для параметров квадратичной модели регрессии, проверить адекватность полученной квадратичной модели с тем же уровнем значимости;

– найти границы доверительных интервалов для параметров квадратичной модели;

– построить графики полученных моделей регрессии и графики отклонений функции регрессии от экспериментальных данных.

4.3. Пример выполнения типового расчета.

4.3.1. Задача 1.

Результаты эксперимента представлены в первых двух столбцах таблицы 4.1. По отдельной серии из n_1=19 экспериментов найдена оценка дисперсии S^2=0,96. Найдем оценки параметров линейной регрессии *Y* на *x*.

Таблица 4.1

Исходные данные и результаты расчета (к задаче 1)

x	Y	$x-\bar{x}$	X	X^2	YX	$Y_{лин}$	ΔY	ΔY^2
5	19,1	–45	–9	81	–171,9	19,65	–0,55	0,3025
20	25,4	–30	–6	36	–152,4	24,6	0,8	0,64
40	33,0	–10	–2	4	–66,0	31,2	1,8	3,24
50	34,1	0	0	0	0	34,5	–0,4	0,16
55	35,0	5	1	1	35,0	36,15	–1,15	1,3225
65	38,0	15	3	9	114,0	39,45	–1,45	2,1025
75	42,4	25	5	25	212,0	42,75	–0,35	0,1225
90	49,0	40	8	64	392,0	47,7	1,3	1,69
$\sum 400$	276	0	0	220	362,7		0	9,58

Сначала найдем решение задачи регрессии в кодированных значениях переменной x (4.14). Введем новую переменную по формуле $X = \dfrac{x-\bar{x}}{h}$, где $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{8}\cdot 400 = 50.$ Значения величины $x-\bar{x}$ являются целыми числами с общим множителем $h = 5$. Поэтому переменная $X = \dfrac{x-50}{5}$ принимает целые значения, не имеющие общего множителя.

По формулам (4.15) находим оценки коэффициентов линейной регрессии

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{276}{8} = 34,5; \qquad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^{\,2}} = \frac{362,7}{220} = 1,6484 \approx 1,65.$$

Получили линейную модель регрессии

$$Y_{лин} = 34,5 + 1,65X. \tag{4.30}$$

Вычисляем значения линейной функции регрессии $\hat{Y}_{лин}$ по формуле (4.30) при всех значениях аргумента X, а затем рассчитываем $\Delta Y_i = \hat{Y}_i - \hat{Y}_{i\,лин}$ отклонения экспериментальных значений $\hat{Y}_i$ от значений $\hat{Y}_{i\,лин}$, полученных по функции регрессии, и, для контроля, суммируем значения ΔY_i. Условие (4.16) выполняется.

Далее суммируем значения ΔY_i^2 для нахождения дисперсии адекватности (4.17). Все расчеты приведены в таблице 4.1.

$$S^2_{\text{ад}} = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{n-2} = \frac{9,58}{6} = 1,597.$$

Проверяем адекватность линейной модели регрессии, используя критерий Фишера (4.11).

$$F = \frac{S^2_{ад}}{S^2_{экс}} = \frac{1,597}{0,96} = 1,664. \quad k_{ад} = n-2.$$

Квантиль распределения Фишера $F_{1-\alpha}(k_{ад}, k_{экс}) = F_{0,95}(6;18) = 2,66.$ Так как

$$F = 1,664 < 2,66 = F_{1-\alpha}(k_{ад}, k_{экс}),$$

то гипотеза об адекватности линейной модели регрессии принимается.

Найдем доверительные интервалы для коэффициентов регрессии. Границы доверительных интервалов для коэффициентов B_1 и B_2 найдем по формулам (4.18). Так как линейная модель регрессии адекватна, в качестве оценки дисперсии возьмем дисперсию адекватности.

$$B_1 = 34,5 \pm \varepsilon_1; \qquad \varepsilon_1 = t_{1-\frac{\alpha}{2}}(k_{ад}) \frac{S_{ад}}{\sqrt{n}} = 2,447 \frac{\sqrt{1,597}}{\sqrt{8}} = 1,093;$$

$$B_2 = 1,65 \pm \varepsilon_2; \qquad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(k_{ад}) \frac{S_{ад}}{\sqrt{\sum_{i=1}^{n} X_i^{\,2}}} = 2,447 \frac{\sqrt{1,597}}{\sqrt{220}} = 0,208.$$

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

$$Y_{лин} = \beta_1 + \beta_2 x = 34,5 + 1.65 \frac{x-50}{5} = 18,0 + 0.33x. \qquad (4.31)$$

Границы доверительных интервалов для коэффициентов β_1 и β_2 (4.18):

$$\beta_1 = 18,0 \pm \tilde{\varepsilon}_1; \qquad \hat{\varepsilon}_1 = t_{1-\frac{\alpha}{2}}(k_{ад})\, S_{ад} \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{h^2 \sum_{i=1}^{n} X_i^{\,2}}} = 2,447\sqrt{1,597}\sqrt{\frac{1}{8} + \frac{2500}{25 \cdot 220}} = 2,355;$$

$$\beta_2 = 0,33 \pm \hat{\varepsilon}_2; \qquad \hat{\varepsilon}_2 = t_{1-\frac{\alpha}{2}}(k_{ад}) \frac{S_{ад}}{h\sqrt{\sum_{i=1}^{n} X_i^{\,2}}} = 2,447 \frac{\sqrt{1,597}}{5\sqrt{220}} = 0,0416.$$

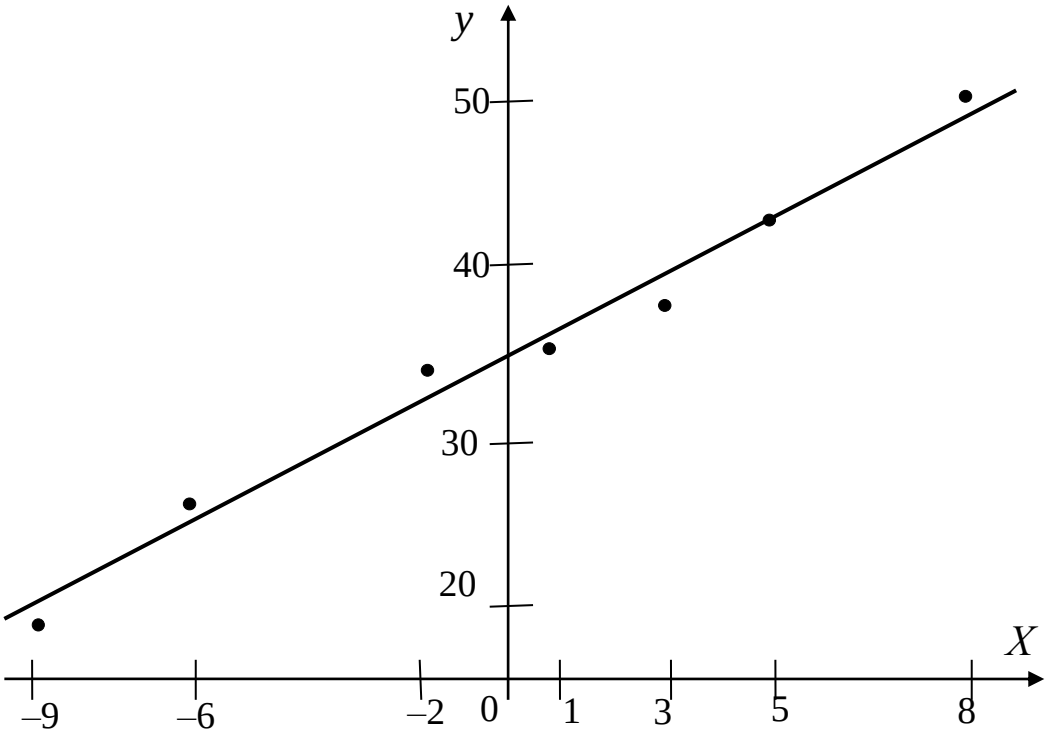


Рис. 4.1. График линейной модели регрессии

На рис. 4.1 построен график линейной модели регрессии. Точками помечены результаты эксперимента. На рисунке 4.2 приведен график отклонений экспериментальных данных от функции регрессии. Отклонения носят хаотический характер, что подтверждает адекватность полученной модели.

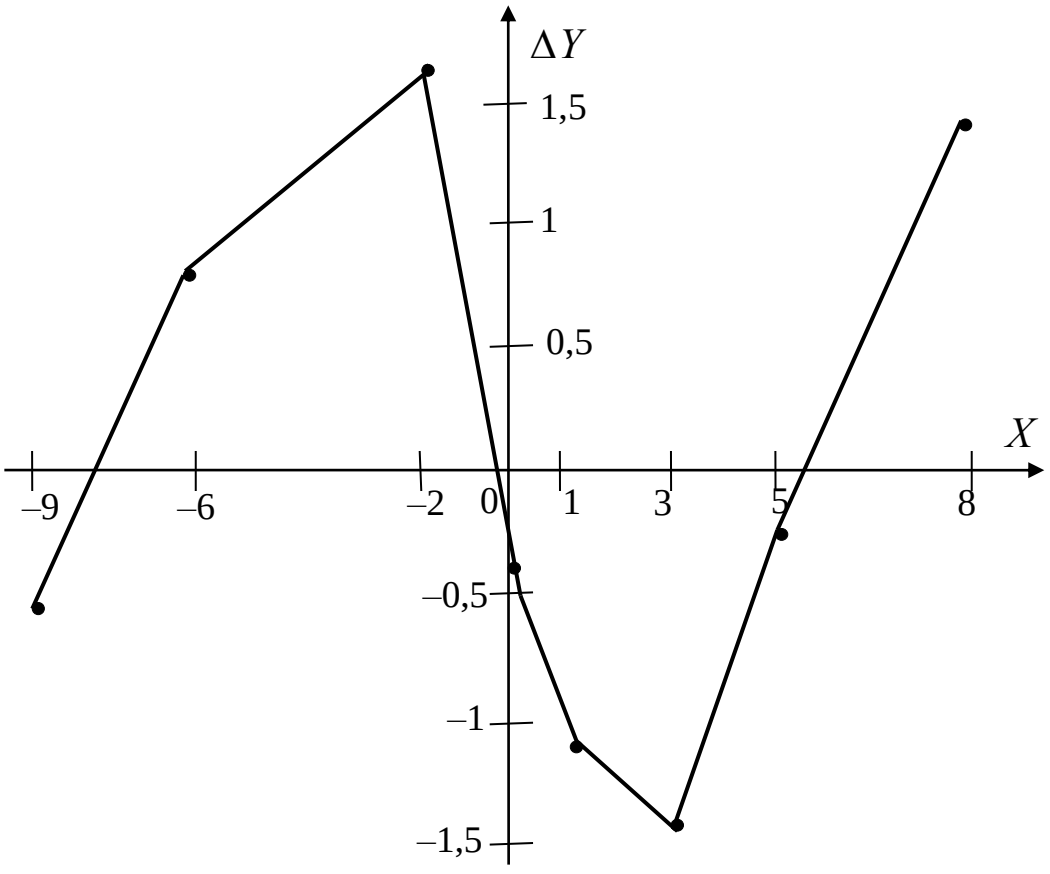


Рис. 4.2. График отклонений экспериментальных данных от линейной функции регрессии.

Выводы по работе. По приведенным исходным данным методом наименьших квадратов построена адекватная линейная модель регрессии с уровнем значимости α = 0,05,

$$S^2_{\text{ад}} = 1,597.$$

Уравнение регрессии в кодированных значениях переменной $X = \frac{x-50}{5}$:

$$y_{лин} = B_1 + B_2 X,$$

где $B_1 = 34,5 \pm 1,093;$ $B_2 = 1,65 \pm 0,208.$

Уравнение регрессии в исходной переменной *x*:

$$y_{лин} = \beta_1 + \beta_2 x,$$

где $\beta_1 = 18,0 \pm 2,355;$ $\beta_2 = 0,33 \pm 0,0416.$

4.3.2. Задача 2.

Результаты эксперимента представлены в первых двух столбцах таблицы 4.2. По отдельной серии из n_2 =20 экспериментов найдена оценка дисперсии S^2 =0,176. Найдем оценки параметров линейной регрессии *Y* на *x*.

Таблица 4.2

Исходные данные и результаты расчета линейной модели регрессии (к задаче 2)

x	Y	$x-\bar{x}$	X	X^2	$\widehat{YX}$	$\widehat{Y}_{лин}$	$\Delta Y_{лин}$	$\Delta Y^2_{лин}$
0	12,28	–0,5	–5	25	38,6	14,923	–2,643	6,9854
0,2	18,66	–0,3	–3	9	4,02	17,785	0,875	0,7656
0,4	22,80	–0,1	–1	1	–2,80	20,647	2,153	4,6354
0,6	24,97	0,1	1	1	4,97	23,509	1,461	2,1345
0,8	26,70	0,3	3	9	20,10	26,371	0,329	0,1082
1,0	27,06	0,5	5	25	35,3	29,233	–2,173	4,7219
$\sum 3,0$	132,67	0	0	70	100,19		–0,002	19,3510

Найдем решение задачи линейной регрессии в кодированных значениях переменной *x* (4.14). Введем новую переменную по формуле $X = \frac{x-\bar{x}}{h}$, где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6} \cdot 3,0 = 0,5.$ Значения величины $x-\bar{x}$ будут целые числа, не имеющие общего множителя, если принять *h* = 0,1. Поэтому введем переменную $X = \frac{x-0,5}{0,1}$.

По формулам (4.15) находим оценки коэффициентов линейной регрессии

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{132,47}{6} = 22,078;, \qquad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^{\,2}} = \frac{100,19}{70} = 1,431.$$

Получили линейную модель регрессии

$$\widehat{Y}_{лин} = 22,078 + 1,431X. \tag{4.32}$$

Вычисляем значения линейной модели регрессии $\widehat{Y}_{лин}$ по формуле (4.32) при всех значениях аргумента *X*, а затем рассчитываем $\Delta Y_{i\,лин} = \widehat{Y}_i - \widehat{Y}_{i\,лин}$ отклонения экспериментальных значений $\widehat{Y}_i$ от значений $\widehat{Y}_{i\,лин}$, полученных по функции регрессии. Сумма $\sum_{i=1}^{n} \Delta Y_{i\,лин}$ не равна нулю в силу того, что при вычислении коэффициентов регрессии результаты округлялись с точностью трех знаков после запятой и при вычислении суммы отклонений набежала ошибка округления. Далее суммируем значения ΔY_i^2 для нахождения дисперсии адекватности (4.17). Все расчеты приведены в таблице 4.2.

$$S^2_{\text{ад лин}} = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{n-2} = \frac{19,351}{4} = 4,838.$$

Проверяем адекватность линейной модели регрессии, используя критерий Фишера (4.11).

$$F_{лин} = \frac{S^2_{ад\,лин}}{S^2_{экс}} = \frac{4,838}{0,176} = 27,487.$$

Квантиль распределения Фишера $F_{1-\alpha}(k_{ад\,лин}, k_{экс}) = F_{0,95}(4;19) = 2,90.$ Так как

$$F_{лин} = 27,487 > 2,90 = F_{1-\alpha}(k_{ад\,лин}, k_{экс}),$$

То гипотеза об адекватности линейной модели регрессии отвергается.

Построим квадратичную модель регрессии с кодированной переменной (4.23)

$$y = B_1 + B_2 X + B_3 X^2 = B_1 + B_2 \frac{x - 0,5}{0,1} + B_3 \left(\frac{x - 0,5}{0,1} \right)^2. \tag{4.33}$$

Заметим, что в нашем случае $\sum_{i=1}^{n} X_i^{\,3} = 0$, поэтому для оценок коэффициентов регрессии можно воспользоваться формулами (4.25). Все расчеты приведены в таблице 4.3.

Таблица 4.3

Результаты расчета квадратичной модели регрессии (к задаче 2)

X^3	YX^2	X^4	$\widehat{Y}_{кв}$	$\Delta Y_{кв}$	$\Delta Y_{кв}^{\,2}$
–125	307,00	626	12,558	–0,278	0,07728

–27	167,94	81	18,258	0,402	0,16160
–1	22,80	1	22,540	0,26	0,06760
1	24,97	1	25,402	–0,432	0,18662
27	240,30	81	26,844	–0,144	0,02074
125	676,50	625	26,878	0,182	0,03312
0	1439,51	1414		–0,010	0, 54996

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} X_i^4 \cdot \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i^2 \cdot \sum_{i=1}^{n} Y_i X_i^2}{n\sum_{i=1}^{n} X_i^4 - \left(\sum_{i=1}^{n} X_i^2\right)^2} = \frac{1414 \cdot 132,47 - 70 \cdot 1439,51}{6 \cdot 1414 - 70^2} = 24,148;$$

$$\tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} = \frac{100,19}{70} = 1,431;$$

$$\tilde{B}_3 = \frac{n\sum_{i=1}^{n} Y_i X_i^2 \cdot - \sum_{i=1}^{n} X_i^2 \cdot \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^4 - \left(\sum_{i=1}^{n} X_i^2\right)^2} = \frac{6 \cdot 1439,51 - 70 \cdot 132,47}{6 \cdot 1414 - 70^2} = -0,1774.$$

Получили квадратичную модель регрессии

$$Y_{кв} = 24,148 + 1,431X - 0,1774X^2. \qquad (4.34)$$

Для расчета дисперсии адекватности (4.28) вычисляем $\Delta Y_{i\,кв} = \widehat{Y}_i - \widehat{Y}_{i\,кв}$ отклонения экспериментальных значений $\widehat{Y}_i$ от значений $\widehat{Y}_{i\,кв}$, полученных по функции регрессии. Отличие $\sum_{i=1}^{n} Y_{i\,кв}$ от нуля объясняется ошибками округления. Проверяем адекватность квадратичной модели регрессии

$$S^2_{\text{ад кв}} = \frac{\sum_{i=1}^{n} (\Delta Y_{i\,кв})^2}{n-3} = 0,1833. \qquad F_{кв} = \frac{S^2_{ад\,кв}}{S^2_{экс}} = \frac{0,1833}{0,176} = 1,042.$$

Квантиль распределения Фишера $F_{1-\alpha}(k_{ад\,кв}, k_{экс}) = F_{0,95}(3;19) = 3,13.$ Так как

$$F_{кв} = 1,042 < 3,13 = F_{1-\alpha}(k_{ад\,кв}, k_{экс}),$$

то гипотеза об адекватности квадратичной модели регрессии принимается.

Найдем доверительные интервалы для коэффициентов регрессии. Границы доверительных интервалов (4.27) для коэффициентов B_1, B_2 и B_3:

$$B_1 = 24{,}148 \pm \varepsilon_1; \qquad \varepsilon_1 = t_{1-\frac{\alpha}{2}}(n-3)\, S_{ад\,кв} \sqrt{\frac{\sum_{i=1}^{n} X_i^{\,4}}{n\sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2}\right)^2}} = 3{,}182\sqrt{0{,}1833}\sqrt{\frac{1414}{6\cdot 1414 - 70^2}} = 0{,}856;$$

$$B_2 = 1{,}431 \pm \varepsilon_2; \qquad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(n-3)\, S_{ад\,кв} \frac{1}{\sqrt{\sum_{i=1}^{n} X_i^{\,2}}} = 3{,}182\sqrt{0{,}1833}\frac{1}{\sqrt{70}} = 0{,}1628;$$

$$B_3 = -0{,}1774 \pm \varepsilon_3; \qquad \varepsilon_3 = t_{1-\frac{\alpha}{2}}(n-3)\, S_{ад\,кв} \sqrt{\frac{n}{n\sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2}\right)^2}} = 3{,}182\sqrt{0{,}1833}\sqrt{\frac{6}{6\cdot 1414 - 70^2}} = 0{,}05574.$$

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

$$Y_{кв} = \beta_1 + \beta_2 x + \beta_3 x^2 = 24{,}148 + 1{,}431\frac{x-0{,}5}{0{,}1} - 0{,}1774\left(\frac{x-0{,}5}{0{,}1}\right)^2 = 12{,}558 + 32{,}05x - 17{,}74x^2.$$

Границы доверительных интервалов для коэффициентов β_1, β_2 и β_3 (4.29):

$$\beta_1 = 12{,}558 \pm \hat{\varepsilon}_1; \qquad \hat{\varepsilon}_1 = t_{1-\frac{\alpha}{2}}(n-3)\, S_{ад\,кв} \sqrt{\frac{\sum_{i=1}^{n} X_i^{\,4} + \frac{\overline{x}^4}{h^4} n}{n\sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2}\right)^2} + \frac{\overline{x}^2}{h^2\sum_{i=1}^{n} X_i^{\,2}}} =$$

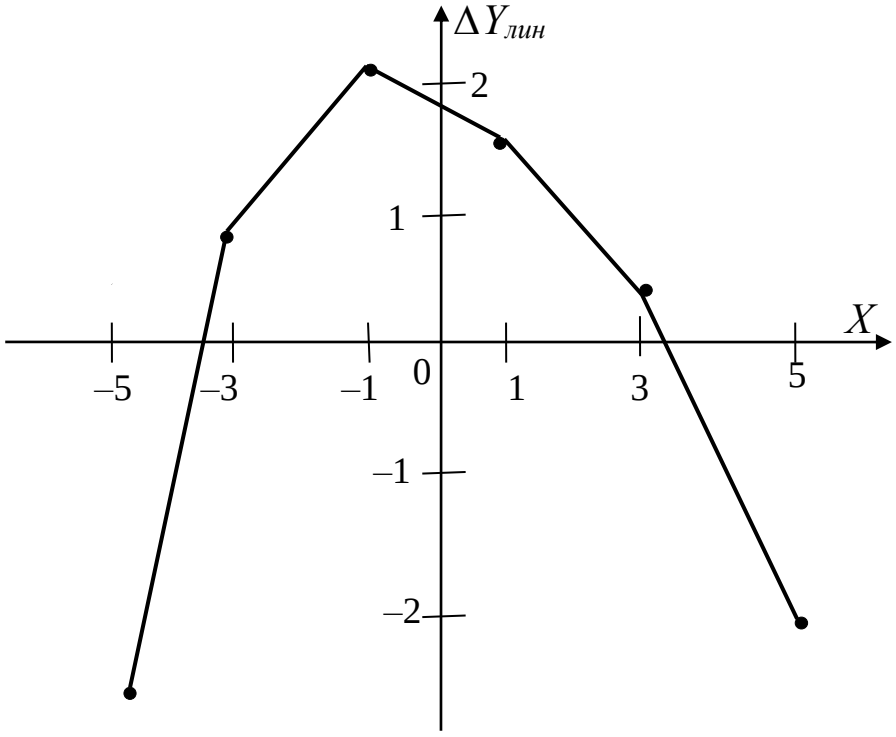
$$= 3{,}182\sqrt{0{,}1833}\sqrt{\frac{1414 + 6\cdot 625}{6\cdot 1414 - 70^2} + \frac{25}{70}} = 1{,}8267;$$

$$\beta_2 = 32{,}05 \pm \hat{\varepsilon}_2; \qquad \hat{\varepsilon}_2 = t_{1-\frac{\alpha}{2}}(n-3)\frac{S_{ад\,кв}}{h} \sqrt{\frac{1}{\sum_{i=1}^{n} X_i^{\,2}} + \frac{4n\overline{x}^2}{h^2\left(n\sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2}\right)^2\right)}} =$$

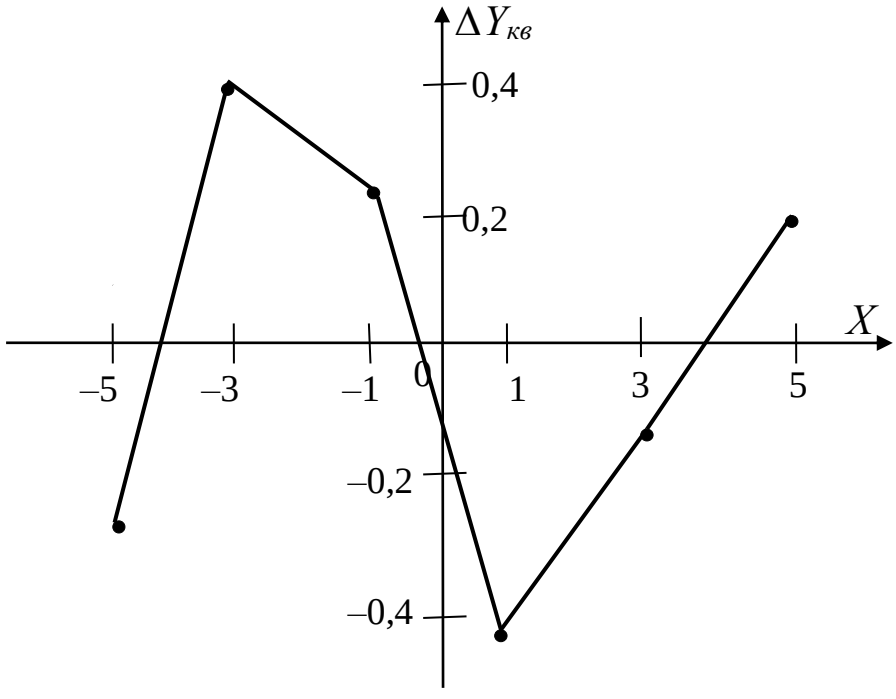
$$= 3{,}182\frac{\sqrt{0{,}1833}}{0{,}1}\sqrt{\frac{1}{70} + \frac{24\cdot 25}{6\cdot 1414 - 70^2}} = 5{,}8069;$$

$$\beta_3 = -17{,}74 \pm \hat{\varepsilon}_3; \qquad \hat{\varepsilon}_3 = t_{1-\frac{\alpha}{2}}(n-3)\frac{S_{ад\,кв}}{h^2} \sqrt{\frac{n}{n\sum_{i=1}^{n} X_i^{\,4} - \left(\sum_{i=1}^{n} X_i^{\,2}\right)^2}} =$$

$$= 3{,}182\frac{\sqrt{0{,}1833}}{0{,}01}\sqrt{\frac{6}{6\cdot 1414 - 70^2}} = 5{,}5740.$$



а)



б)

Рис. 4.3. График отклонения экспериментальных данных от линейной модели (*а*); от квадратичной модели (*б*).

Для графического анализа результатов расчета построены графики отклонений линейной и квадратичной моделей регрессии от экспериментальных данных.

На рис. 4.3 *а*) представлен график отклонений $\Delta Y_{лин}$ по приведенной выше таблице 4.2. Из рис. 4. 3 *а*) видна не только непригодность линейной модели (что следует уже из больших значений отклонений $\Delta Y_{лин}$, явно превышающих погрешность эксперимента, но и

целесообразность расчета квадратичной модели, так как расположение точек наводит на мысль о параболе.

На рис. 4.3 *b*) представлен график отклонений $\Delta Y_{\text{кв}}$ по таблице 4.3, но построенный уже в другом масштабе с увеличением в 5 раз. Из рис. 4.3 *b*) видно, что отклонения от параболы, то есть $|\Delta Y_{\text{кв}}|$ малы (имеют порядок ошибок эксперимента); это свидетельствует о соответствии квадратичной модели регрессии результатам эксперимента.

Надо также графически сравнить линейную и квадратичную модели с экспериментальными точками. На рис. 4.4 такое сравнение проведено для рассматриваемого примера; оно показывает соответствие квадратичной модели с экспериментом.

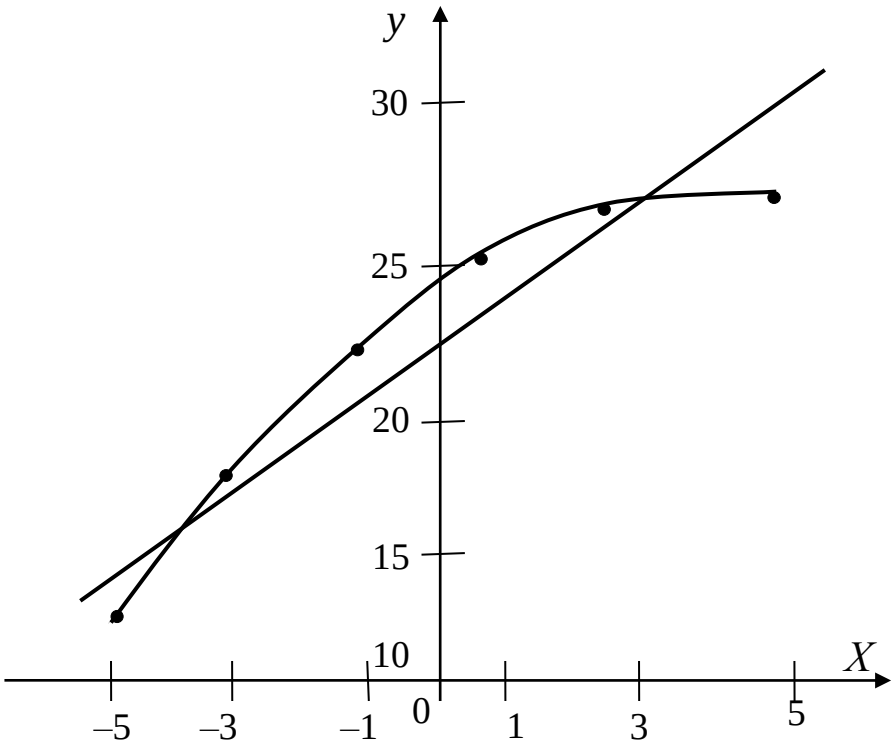


Рис. 4.4. Сравнение линейной и квадратичной моделей с данными эксперимента.

Выводы по работе. По приведенным исходным данным построенная методом наименьших квадратов линейная модель регрессии неадекватна, квадратичная модель адекватна (уровень значимости $\alpha = 0{,}05$).

$$S^2_{\text{ад кв}} = 0{,}1833\,.$$

Уравнение регрессии в кодированных значениях переменной $X = \dfrac{x - 0{,}5}{0{,}1}$:

$$y_{кв} = B_1 + B_2 X + B_3 X^2$$

где $B_1 = 24{,}148 \pm 0{,}856;$ $B_2 = 1{,}431 \pm 0{,}1628;$ $B_3 = -0{,}1774 \pm 0{,}05574.$

Уравнение регрессии в исходной переменной *x*:

$$y_{кв} = \beta_1 + \beta_2 x + \beta_3 x^2,$$

где $\beta_1 = 12,558 \pm 1,8267$; $\beta_2 = 32,05 \pm 5,8069$; $\beta_3 = -17,74 \pm 5,5740$.