

**Галанов Ю.И.****Лабораторный практикум по мат. статистике****Классический регрессионный анализ**

Задание.

1. Смоделировать регрессионную зависимость в виде многочлена третьей степени.
2. Смоделировать независимую выборку для оценки дисперсии при фиксированном значении независимой переменной.
3. Оценить параметры полиномиальной модели методом наименьших квадратов.
4. Проверить гипотезу об адекватности модели с помощью критерия Фишера.
5. Оценить коэффициент множественной корреляции (детерминации) и проверить гипотезу об его значимости.
6. Проверить гипотезу о значимости коэффициентов модели.
7. Исследовать остатки.
8. Построить доверительные интервалы для предсказанных значений.

Указание.

1. Проведите проверку адекватности моделей, начиная с многочлена первой степени до многочлена 4-й степени включительно. Для этого подключайте необходимое число регрессоров в матрице F.
2. Проведите анализ остатков.
3. Сформулируйте мотивированные выводы об адекватности моделей.
4. Проверку гипотезы о значимости коэффициентов модель проведите при больших и малых значениях объема выборки.
5. Исключить незначимые коэффициенты.

Моделируем данные

$$b_0 := 100 \quad b_1 := 6 \quad b_2 := -40 \quad b_3 := 5$$

$$H(x) := b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$$

$$w(x, y) := \sin(2 \cdot \pi \cdot x) \cdot \sqrt{-2 \cdot \ln(y)}$$

$$s := 1000$$

$$aL := -5$$

$$aR := 10$$

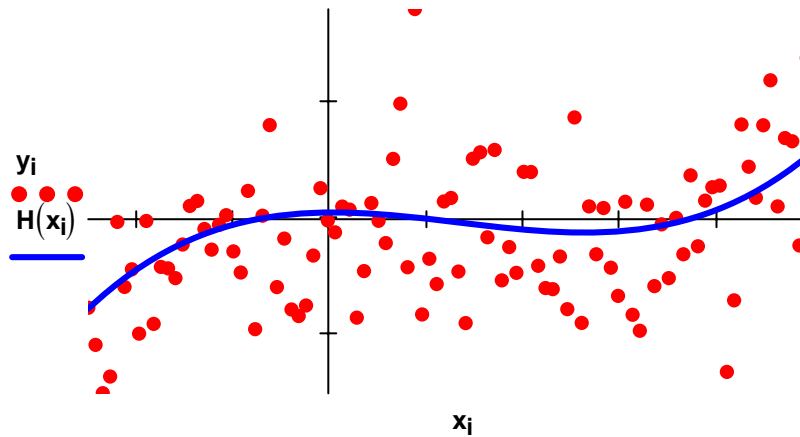
$$N := 100 \quad h := \frac{aR - aL}{N}$$

$$i := 0.. N - 1$$

$$x_i := aL + i \cdot h$$

Наблюдаемые значения =>

$$y_i := H(x_i) + s \cdot w(\text{rnd}(1), \text{rnd}(1))$$



Моделируем вспомогательную выборку

$$n := \text{round}\left(\frac{N}{2}\right) \quad n = 50$$

$$u := 0.. n - 1$$

$$V_u := H(0) + s \cdot w(\text{rnd}(1), \text{rnd}(1))$$

Получим независимую оценку дисперсии

$$sN := \text{Stdev}(V)$$

$$sN = 1.212 \times 10^3$$

Оценка параметров модели

Составляем матрицу регрессоров

$$F_{i,0} := 1$$

$$F_{i,1} := (x_i)$$

$$F_{i,2} := (x_i)^2$$

$$F_{i,3} := (x_i)^3$$

$$F_{i,4} := (x_i)^4$$

Информационная матрица

$$\mathbf{G} := \mathbf{F}^T \cdot \mathbf{F}$$

$$\mathbf{G} = \begin{pmatrix} 100 & 242.5 & 2.463 \times 10^3 & 1.507 \times 10^4 \\ 242.5 & 2.463 \times 10^3 & 1.507 \times 10^4 & 1.329 \times 10^5 \\ 2.463 \times 10^3 & 1.507 \times 10^4 & 1.329 \times 10^5 & 1.043 \times 10^6 \\ 1.507 \times 10^4 & 1.329 \times 10^5 & 1.043 \times 10^6 & 9.114 \times 10^6 \end{pmatrix}$$

Матрица ошибок

$$\mathbf{C} := \mathbf{G}^{-1}$$

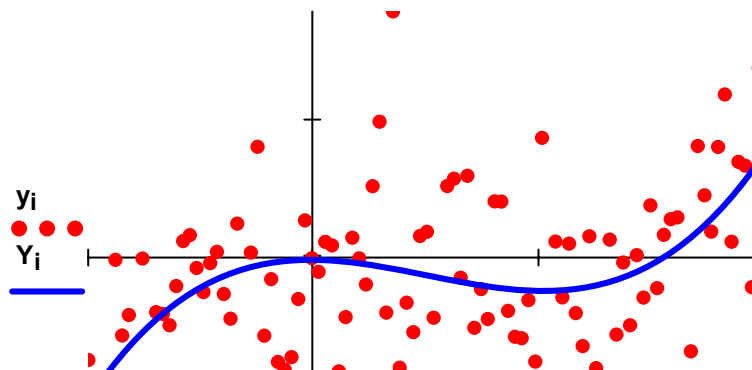
$$\mathbf{C} = \begin{pmatrix} 0.03 & -1.751 \times 10^{-3} & -1.668 \times 10^{-3} & 1.663 \times 10^{-4} \\ -1.751 \times 10^{-3} & 2.008 \times 10^{-3} & 1.158 \times 10^{-4} & -3.963 \times 10^{-5} \\ -1.668 \times 10^{-3} & 1.158 \times 10^{-4} & 1.659 \times 10^{-4} & -1.791 \times 10^{-5} \\ 1.663 \times 10^{-4} & -3.963 \times 10^{-5} & -1.791 \times 10^{-5} & 2.462 \times 10^{-6} \end{pmatrix}$$

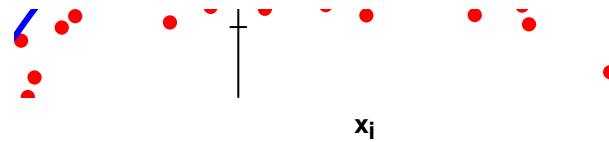
Оценка параметров =>

$$\mathbf{B} := \mathbf{C} \cdot \mathbf{F}^T \cdot \mathbf{y} \quad \mathbf{B} = \begin{pmatrix} -65.713 \\ -7.104 \\ -50.428 \\ 6.709 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 100 \\ 6 \\ -40 \\ 5 \end{pmatrix}$$

Предсказанные значения =>

$$\mathbf{Y} := \mathbf{F} \cdot \mathbf{B}$$





Статистический анализ регрессионной модели

Число параметров модели

$$k := \text{rows}(B)$$

$$k = 4$$

Средние

$$y_{\text{cp}} := \text{mean}(y)$$

$$Y_{\text{cp}} := \text{mean}(Y)$$

$$y_{\text{cp}} = -314.202$$

$$Y_{\text{cp}} = -314.202$$

Остатки

$$e_i := y_i - Y_i$$

Остаточная сумма квадратов

$$Q_{\text{ост}} := \sum_i (e_i)^2$$

$$Q_{\text{ост}} = 1.068 \times 10^8$$

Число степеней свободы

$$r := N - k$$

$$r = 96$$

**Оценка дисперсии
(остаточная дисперсия)**

$$d := \frac{Q_{\text{ост}}}{r}$$

$$d = 1.112 \times 10^6$$

Сумма квадратов, обусловленная уравнением регрессии

$$Q_R := \sum_i (Y_i - Y_{\text{cp}})^2$$

$$Q_R = 3.459 \times 10^7$$

Дисперсия, обусловленная регрессией

$$D_R := \frac{Q_R}{k - 1}$$

$$D_R = 1.153 \times 10^7$$

Полная сумма квадратов

$$Q := \sum_i (y_i - y_{\text{cp}})^2$$

$$Q = 1.414 \times 10^8$$

Проверка

$$Q_R + Q_{\text{ост}} = 1.414 \times 10^8$$

Проверка адекватности модели**Независимая оценка дисперсии**

$$sN^2 = 1.47 \times 10^6$$

**Дисперсионное
отношение Фишера**

$$F := \frac{d}{sN^2}$$

$$F = 0.757$$

$$L := \text{if}\left(F > 1, F, \frac{1}{F}\right)$$

$$L = 1.321$$

Критическая точка

$$F_c := qF(0.95, N - k, n - 1)$$

$$F_c = 1.533$$

**Функция принятия
решения**

$$f := \text{if}(L \leq F_c, \text{"ADEQU"}, \text{"NO_ADEQU"})$$

$$f = \text{"ADEQU"}$$

**Проверка гипотезы о значимости коэффициента
детерминации**

Коэффициент детерминации (множественной корреляции) в случае парной регрессии (функция одной переменной) равен квадрату коэффициента корреляции между наблюдаемыми и предсказанными значениями определяемой переменной. Он является обобщенной (суммарной) мерой качества регрессионной модели.

$$R := \text{corr}(y, Y)$$

$$R = 0.495$$

$$R^2 = 0.245$$

$$\frac{Q_R}{Q} = 0.245$$

Статистика критерия -- отношение дисперсии, обусловленной регрессией к остаточной дисперсии:

$$F := \frac{R^2 \cdot (N - k)}{(1 - R^2) \cdot (k - 1)} \quad F = 10.365 \quad \text{или} \quad \frac{D_R}{d} = 10.365$$

$$L := \text{if} \left(F > 1, F, \frac{1}{F} \right) \quad L = 10.365$$

Критическая точка $F_t := qF(0.95, k - 1, N - k) \quad F_t = 2.699$

Функция принятия решения $f := \text{if}(L < F_t, \text{"NE_ZNATHIM"}, \text{"ZNATHIM"})$

$$f = \text{"ZNATHIM"}$$

Проверка гипотезы о значимости коэффициентов модели

$$m := 0.. \text{rows}(\mathbf{C}) - 1 \quad l := 0.. \text{rows}(\mathbf{C}) - 1$$

Дисперсии оценок

$$SB_m := \sqrt{C_{m,m} \cdot d} \quad SB^T = (183.485 \quad 47.266 \quad 13.583 \quad 1.655)$$

Задаем уровень значимости и находим критическую точку

$$\alpha := 0.05 \quad t_{kr} := qt\left(1 - \frac{\alpha}{2}, N - k\right) \quad t_{kr} = 1.985$$

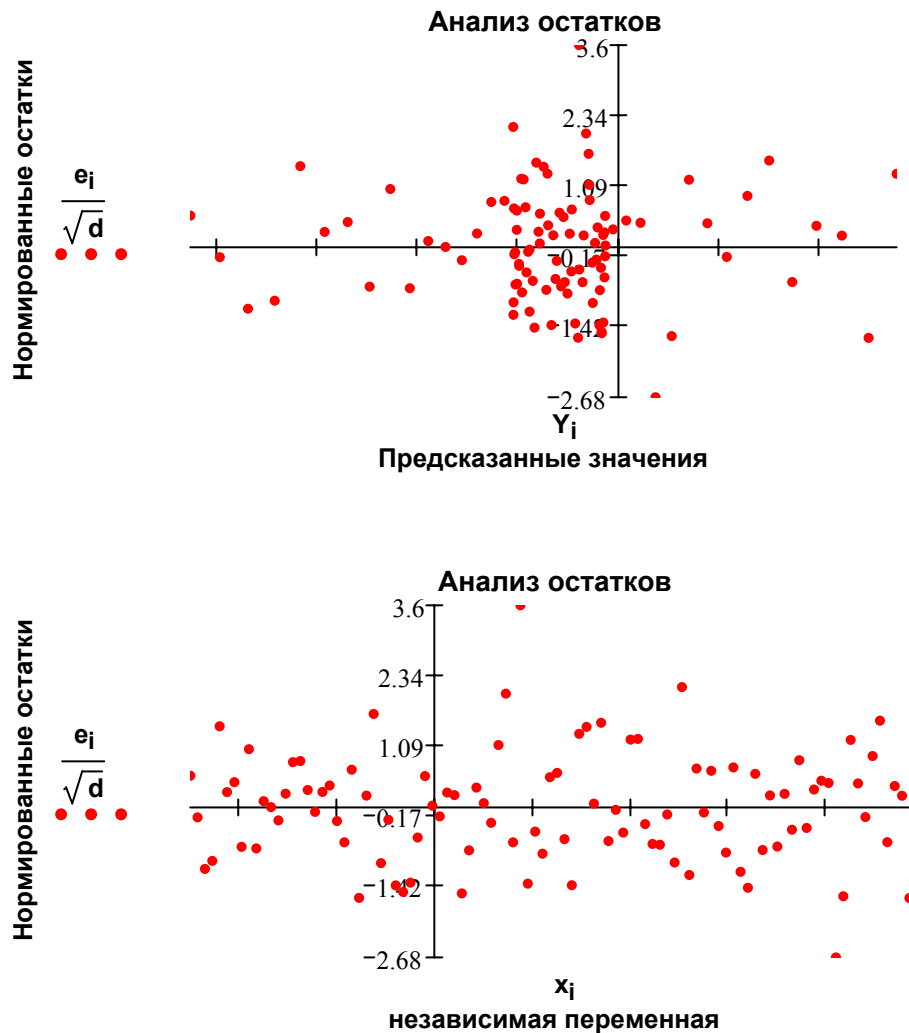
Рассчитываем статистику критерия для каждого коэффициента модели

$$T_m := \frac{|B_m|}{SB_m} \quad T^T = (0.358 \quad 0.15 \quad 3.713 \quad 4.054)$$

Вводим функцию принятия решения и находим ее значение, на основании которой делаем выводы о значимости коэффициентов. Если коэффициент незначим, то соответствующий член в модели надо исключить.

$$Kr_m := \text{if}(T_m < t_{kr}, \text{"NeZnathin"}, \text{"Znathim"}) \quad Kr = \begin{pmatrix} \text{"NeZnathin"} \\ \text{"NeZnathin"} \\ \text{"Znathim"} \\ \text{"Znathim"} \end{pmatrix}$$

Анализ остатков



Корреляционная матрица коэффициентов модели

$$K_{l,m} := \frac{C_{l,m}}{\sqrt{C_{l,l} \cdot C_{m,m}}}$$

$$K = \begin{pmatrix} 1 & -0.225 & -0.744 & 0.609 \\ -0.225 & 1 & 0.201 & -0.564 \\ -0.744 & 0.201 & 1 & -0.886 \\ 0.609 & -0.564 & -0.886 & 1 \end{pmatrix}$$

Доверительные интервалы для предсказанного значения

$$\mathbf{f}(\mathbf{x}) := \begin{bmatrix} 1 \\ \mathbf{x} \\ (\mathbf{x})^2 \\ (\mathbf{x})^3 \\ \mathbf{x}^4 \end{bmatrix}$$

$$\mathbf{f}(\mathbf{x}) := \begin{pmatrix} 1 \\ \mathbf{x} \\ \mathbf{x}^2 \\ \mathbf{x}^3 \end{pmatrix}$$

$$Y_{2i} := Y_i + t_{kr} \cdot \sqrt{\left(\mathbf{d} \cdot (\mathbf{f}(\mathbf{x}_i))^T \cdot \mathbf{C} \cdot \mathbf{f}(\mathbf{x}_i) \right)_{0,0}}$$

$$Y_{1i} := Y_i - t_{kr} \cdot \sqrt{\left(\mathbf{d} \cdot (\mathbf{f}(\mathbf{x}_i))^T \cdot \mathbf{C} \cdot \mathbf{f}(\mathbf{x}_i) \right)_{0,0}}$$

$$\sqrt{\mathbf{d}} = 1.055 \times 10^3$$

