Решение задач по теме: метод наименьших квадратов
(обработка экспериментальных данных)

Задание.

Проводится анализ взаимосвязи количества населения (X) и количества практикующих врачей (Y) в регионе.

Годы	81	82	83	84	85	86	87	88	89	90
X, млн. чел.	10	10,3	10,4	10,55	10,6	10,7	10,75	10,9	10,9	11
Y, тыс. чел.	12,1	12,6	13	13,8	14,9	16	18	20	21	22

Оцените по МНК коэффициенты линейного уравнения регрессии $y = b_0 + b_1 x$.

Существенно ли отличаются от нуля найденные коэффициенты?

Проверьте значимость полученного уравнения при α = 0,01.

Если количество населения в 1995 году составит 11,5 млн. чел., каково ожидаемое количество врачей? Рассчитайте 99%-й доверительный интервал для данного прогноза. Рассчитайте коэффициент детерминации

Решение.

Оценим по МНК коэффициенты линейного уравнения регрессии $y = b_0 + b_1 x$. Параметры b_0, b_1 по методу наименьших квадратов можно найти из системы нормальных уравнений:

$$\begin{cases} b_1 \sum x_i^2 + b_0 \sum x_i = \sum x_i y_i \\ b_1 \sum x_i + b_0 n = \sum y_i \end{cases}$$

где суммирование ведется по i от 1 до n, n = 10. Составим расчетную таблицу:

	x_i	y_i	x_i^2	$x_i y_i$
81	10	12,1	100	121
82	10,3	12,6	106,09	129,78
83	10,4	13	108,16	135,2
84	10,55	13,8	111,3025	145,59
85	10,6	14,9	112,36	157,94
86	10,7	16	114,49	171,2
87	10,75	18	115,5625	193,5
88	10,9	20	118,81	218
89	10,9	21	118,81	228,9
90	11	22	121	242
Сумма	**106,1**	**163,4**	**1126,585**	**1743,11**

Получаем систему:

$$\begin{cases} 1126,585 b_1 + 106,1 b_0 = 1743,11, \\ 106,1 b_1 + 10 b_0 = 163,4; \end{cases}$$

откуда находим $b_0 = -99,535$, $b_1 = 10,921$, то есть получаем функцию:

$\hat{y}_x = y = 10,921x - 99,535$.

Выясним, существенно ли отличаются от нуля найденные коэффициенты.
Определим значимость коэффициентов регрессии, сопоставив стандартные ошибки с величинами самих коэффициентов:

$t_{b1} = \dfrac{b_1}{m_{b1}}$; $\qquad t_{b0} = \dfrac{b_0}{m_{b0}}$

$m_{b0} = S_{\text{ост}} \cdot \sqrt{\dfrac{\sum x^2}{n \cdot \sum (x - \bar{x})^2}}$; $\qquad m_{b1} = S_{\text{ост}} \cdot \sqrt{\dfrac{1}{\sum (x - \bar{x})^2}}$,

где $S_{\text{ост}} = \sqrt{\dfrac{\sum (y - \hat{y}_x)^2}{n - p - 1}}$

На основании данных расчётной таблицы найдем остаточное среднее квадратическое отклонение: $S_{\text{ост}} = \sqrt{\dfrac{18,611}{10 - 1 - 1}} \approx 1,525$

Величины стандартных ошибок для параметров b_0, b_1 равны:

$m_{b0} = 1,525 \cdot \sqrt{\dfrac{1126,585}{10 \cdot 0,864}} = 17,417$;

$m_{b1} = 1,525 \cdot \sqrt{\dfrac{1}{0,864}} = 1,641$.

Из приведённого выше расчётные значения t-критерия Стьюдента можно определить следующим образом:

$t_{b1} = \dfrac{b_1}{m_{b1}} = \dfrac{10,921}{1,641} = 6,656$; $\quad t_{b0} = \dfrac{b_0}{m_{b0}} = \dfrac{-99,535}{17,417} = -5,715$.

Табличное значение t-критерия при α = 0,01 равно $t_{\text{ò àáë}}(0,01;8) = 3,36$. Таким образом, все коэффициенты регрессии оказались значимы (существенно отличны от нуля), так как расчётные значения t-критерия выше табличного значения (по модулю).

Проверим значимость полученного уравнения при α = 0,01. Рассчитаем эмпирическую величину F-критерия Фишера:

$F_p = \dfrac{R^2}{1 - R^2} * \dfrac{n - p - 1}{p}$.

Определим величину R^2:

$R^2 = 1 - \dfrac{\sum (y - \hat{y}_x)^2}{\sum (y - \bar{y})^2} = 1 - \dfrac{18,611}{121,664} = 0,847$

Тогда:

$F_p = \dfrac{0,847}{1 - 0,847} * \dfrac{10 - 1 - 1}{1} = 44,298$

По таблице F-распределения Снедекора-Фишера при α = 0,01 и $К_1$ = 1, $К_2$ = 10 – 2 = 8 величина $F_т$ = 11,26. Это означает, что гипотеза H_0 о несущественности связи между Y и X с вероятностью ошибочности суждения α = 0,01 отклоняется, то есть связь между этими переменными может быть признана существенной. Уравнение значимо в целом.

Если количество населения в 1995 году составит 11,5 млн. чел., то ожидаемое количество врачей будет $\hat{y}_x = y = 10,921 \cdot 11,5 - 99,535 = 26,057$ тыс. человек.

Рассчитаем 99%-й доверительный интервал для данного прогноза при $x_p = 11,5$. Используем формулу: $\hat{y}_{xp} - dy \le \hat{y}_{xp} \le \hat{y}_{xp} + dy$, где

$$dy = t_{\text{табл}} \cdot S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} = 3,36 \cdot 1,525 \cdot \sqrt{1 + \frac{1}{10} + \frac{(11,5 - 10,61)^2}{0,864}} = 7,277$$

Тогда:

$$26,057 - 7,277 \le \hat{y}_{xp} \le 26,057 + 7,277,$$

$$18,78 \le \hat{y}_{xp} \le 33,334.$$

Приведём расчётную таблицу:

год	x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i$	$\left(y_i - \hat{y}_i\right)^2$	$\left(x_i - \bar{x}\right)^2$	$\left(y_i - \bar{y}\right)^2$
81	10	12,1	100	121	9,675	5,880625	0,3721	17,9776
82	10,3	12,6	106,09	129,78	12,9513	0,123412	0,0961	13,9876
83	10,4	13	108,16	135,2	14,0434	1,088684	0,0441	11,1556
84	10,55	13,8	111,3025	145,59	15,68155	3,54023	0,0036	6,4516
85	10,6	14,9	112,36	157,94	16,2276	1,762522	1E-04	2,0736
86	10,7	16	114,49	171,2	17,3197	1,741608	0,0081	0,1156
87	10,75	18	115,5625	193,5	17,86575	0,018023	0,0196	2,7556
88	10,9	20	118,81	218	19,5039	0,246115	0,0841	13,3956
89	10,9	21	118,81	228,9	19,5039	2,238315	0,0841	21,7156
90	11	22	121	242	20,596	1,971216	0,1521	32,0356
Сумма	**106,1**	**163,4**	**1126,585**	**1743,11**	**163,3681**	**18,61075**	**0,864**	**121,664**