Задача. На основе данных о доходах Y, расходах на промышленные товары X_2, наличии детей (табл. 1), необходимо построить модель с фиктивной переменной D (принять D = 1, сли дети есть; D = 0 при их отсутствии), вида:

$$\hat{Y} = b_0 + b_2 X_2 + b \cdot D .$$

Проверить статистическую значимость коэффициентов. Сделать выводы.

Таблица 1

Сведения о доходах, расходах на промышленные товары, о наличии детей

Y	Вариант 4	
	X_2	дети
91,76	3,03	есть
38,68	7,81	нет
34,14	1,63	нет
30,77	4,49	нет
50,02	0,43	нет
34,33	6,31	есть
42,63	5,05	есть
63,47	3,94	нет
19,86	0,29	нет
58,87	5,37	нет
72,45	6,54	есть
29,70	0,93	нет
93,74	1,82	есть
17,77	0,51	есть
78,84	15,87	нет
39,73	1,78	нет
93,87	25,53	нет
86,15	31,97	есть
25,95	2,28	есть
36,95	10,92	есть
45,78	12,76	есть
12,36	0,05	есть

Решение

Рассмотрим уравнение регрессии зависимости доходов Y от величины расходов на промышленные товары X_2 и наличия детей D.

Переменная D – это фиктивная переменная. Она принимает такие значения:

$$D = \begin{cases} 1, \textit{ если есть дети}; \\ 0, \textit{ если нет детей.} \end{cases}$$

Модель регрессии имеет вид:

$$\hat{Y} = b_0 + b_2 X_2 + b \cdot D .$$

Для оценки параметров данного уравнения применяем метод наименьших квадратов. Система нормальных уравнений имеет вид:

$$\begin{cases} \sum Y = n \cdot b_0 + b_2 \cdot \sum X_2 + b \cdot \sum D, \\ \sum Y \cdot X_2 = b_0 \cdot \sum X_2 + b_2 \sum X_2^2 + b \cdot \sum X_2 \cdot D, \\ \sum Y \cdot D = b_0 \cdot \sum D + b_2 \sum X_2 \cdot D + b \cdot \sum D^2. \end{cases}$$

Ввиду того, что D принимает только два значения – 1 и 0, $\sum D = n_1$ – количество семей, в которых есть дети.

Выполним необходимые расчеты:

№ п/п	Y	X_2	D	$Y \cdot X_2$	X_2^2	$X_2 \cdot D$	$Y \cdot D$	D^2
1	91,76	3,03	1	278,03	9,18	3,03	91,76	1
2	38,68	7,81	0	302,09	61,00	0	0	0
3	34,14	1,63	0	55,65	2,66	0	0	0
4	30,77	4,49	0	138,16	20,16	0	0	0
5	50,02	0,43	0	21,51	0,18	0	0	0
6	34,33	6,31	1	216,62	39,82	6,31	34,33	1
7	42,63	5,05	1	215,28	25,50	5,05	42,63	1
8	63,47	3,94	0	250,07	15,52	0	0	0
9	19,86	0,29	0	5,76	0,08	0	0	0
10	58,87	5,37	0	316,13	28,84	0	0	0
11	72,45	6,54	1	473,82	42,77	6,54	72,45	1
12	29,70	0,93	0	27,62	0,86	0	0	0
13	93,74	1,82	1	170,61	3,31	1,82	93,74	1
14	17,77	0,51	1	9,06	0,26	0,51	17,77	1
15	78,84	15,87	0	1251,19	251,86	0	0	0

16	39,73	1,78	0	70,72	3,17	0	0	0
17	93,87	25,53	0	2396,50	651,78	0	0	0
18	86,15	31,97	1	2754,22	1022,08	31,97	86,15	1
19	25,95	2,28	1	59,17	5,20	2,28	25,95	1
20	36,95	10,92	1	403,49	119,25	10,92	36,95	1
21	45,78	12,76	1	584,15	162,82	12,76	45,78	1
22	12,36	0,05	1	0,62	0,00	0,05	12,36	1
∑	1097,82	149,31	11	10000,48	2466,30	81,24	559,87	11

В результате получаем систему:

$$\begin{cases} 1097{,}82 = 22 \cdot b_0 + 149{,}31 \cdot b_2 + 11 \cdot b, \\ 10000{,}48 = 149{,}31 \cdot b_0 + 2466{,}30 \cdot b_2 + 81{,}24 \cdot b, \\ 559{,}87 = 11 \cdot b_0 + 81{,}24 \cdot b_2 + 11 \cdot b. \end{cases}$$

Решив данную систему, находим оценки параметров:

$$b_0 = 38{,}042; b_2 = 1{,}755; b = -0{,}109.$$

Таким образом, уравнение регрессии:

$$\hat{Y} = 38{,}042 + 1{,}755 \cdot X_2 - 1{,}109 \cdot D.$$

Проверим статистическую значимость оценок параметров уравнения регрессии при уровне значимости $\alpha = 0{,}05$.

Сформулируем нулевую (основную) и альтернативную гипотезы:

$$\begin{matrix} H_0 : b_0 = 0; \\ H_a : b_0 \neq 0 \end{matrix}; \quad \begin{matrix} H_0 : b_2 = 0; \\ H_a : b_2 \neq 0 \end{matrix}; \quad \begin{matrix} H_0 : b = 0; \\ H_a : b \neq 0 \end{matrix}$$

При проверке статистической значимости оценок параметров регрессии используется t-критерий Стьюдента.

При этом для каждого фактора используется формула:

$$t_{b_i} = \left| \frac{b_i}{m_{b_i}} \right|,$$

где b_i – коэффициент чистой регрессии при факторе x_i, m_{b_i} – средняя квадратическая (стандартная) ошибка коэффициента регрессии b_i.

Для уравнения множественной регрессии средняя квадратическая ошибка коэффициента регрессии может быть определена по следующей формуле:

$$m_{b_i} = \frac{s_y \sqrt{1 - R^2_{yx_1x_2...x_m}}}{s_{x_i} \sqrt{1 - R^2_{x_1x_2...x_m}}} \cdot \frac{1}{\sqrt{n - m - 1}},$$

где s_y – среднее квадратическое отклонение для признака y; s_{x_i} – среднее квадратическое отклонение для признака x_i; $R^2_{yx_1x_2...x_m}$ – коэффициент детерминации для уравнения множественной регрессии; $R^2_{x_1x_2...x_m}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии; n – m − 1 – число степеней свободы для остаточной суммы квадратов отклонений.

Также, среднюю квадратическую ошибку коэффициентов регрессии можно определить по следующей формуле:

$$m_{b_i} = \sqrt{s_e^2 c_{ii}},$$

где $s_e^2 = \frac{\sum e_i^2}{n - m - 1}$ – остаточная дисперсия; n – количество наблюдений; m – количество объясняющих переменных модели; c_{ii} – диагональний элемент матрицы $(X'X)^{-1}$, $i = \overline{1; p}$. X – матрица факторов с вспомогательной переменной X_0=1, которая вводится в модель, чтобы количество переменных совпадало с количеством параметров.

Найдем остаточную дисперсию и матрицу $(X'X)^{-1}$.

№ п/п	Y	X_2	D	$\hat{Y}$	$e = Y - \hat{Y}$	e^2
1	91,76	3,03	1	43,25	48,51	2353,04
2	38,68	7,81	0	51,75	-13,07	170,86
3	34,14	1,63	0	40,90	-6,76	45,74
4	30,77	4,49	0	45,92	-15,15	229,63
5	50,02	0,43	0	38,80	11,22	125,96
6	34,33	6,31	1	49,01	-14,68	215,49

7	42,63	5,05	1	46,80	-4,17	17,37
8	63,47	3,94	0	44,96	18,51	342,69
9	19,86	0,29	0	38,55	-18,69	349,36
10	58,87	5,37	0	47,47	11,40	130,00
11	72,45	6,54	1	49,41	23,04	530,69
12	29,70	0,93	0	39,67	-9,97	99,49
13	93,74	1,82	1	41,13	52,61	2768,04
14	17,77	0,51	1	38,83	-21,06	443,45
15	78,84	15,87	0	65,90	12,94	167,45
16	39,73	1,78	0	41,17	-1,44	2,06
17	93,87	25,53	0	82,86	11,01	121,30
18	86,15	31,97	1	94,05	-7,90	62,45
19	25,95	2,28	1	41,94	-15,99	255,53
20	36,95	10,92	1	57,10	-20,15	406,09
21	45,78	12,76	1	60,33	-14,55	211,75
22	12,36	0,05	1	38,02	-25,66	658,48
∑	–	–	–	–	–	9706,92

$$s_e^2 = \frac{9706{,}92}{22-2-1} = 510{,}89.$$

$$X = \begin{vmatrix} 1 & 3{,}03 & 1 \\ 1 & 7{,}81 & 0 \\ 1 & 1{,}63 & 0 \\ 1 & 4{,}49 & 0 \\ 1 & 0{,}43 & 0 \\ 1 & 6{,}31 & 1 \\ 1 & 5{,}05 & 1 \\ 1 & 3{,}94 & 0 \\ 1 & 0{,}29 & 0 \\ 1 & 5{,}37 & 0 \\ 1 & 6{,}54 & 1 \\ 1 & 0{,}93 & 0 \\ 1 & 1{,}82 & 1 \\ 1 & 0{,}51 & 1 \\ 1 & 15{,}87 & 0 \\ 1 & 1{,}78 & 0 \\ 1 & 25{,}53 & 0 \\ 1 & 31{,}97 & 1 \\ 1 & 2{,}28 & 1 \\ 1 & 10{,}92 & 1 \\ 1 & 12{,}76 & 1 \\ 1 & 0{,}05 & 1 \end{vmatrix}$$

$$\begin{vmatrix} 0{,}117408 & -0{,}004282 & -0{,}085782 \end{vmatrix}$$

$$(X'X)^{-1} = \begin{vmatrix} \text{-0,004282} & \text{0,000692} & \text{-0,000829} \\ \text{-0,085782} & \text{-0,000829} & \text{0,182810} \end{vmatrix}$$

Средние квадратические ошибки коэффициентов регрессии:

$$m_{b_0} = \sqrt{s_e^2 c_{11}} = \sqrt{510{,}89 \cdot 0{,}117408} = 7{,}745,$$

$$m_{b_2} = \sqrt{s_e^2 c_{22}} = \sqrt{510{,}89 \cdot 0{,}000692} = 0{,}595,$$

$$m_b = \sqrt{s_e^2 c_{33}} = \sqrt{510{,}89 \cdot 0{,}182810} = 9{,}664.$$

Тогда, *t*-статистики равны:

$$t_{b_0} = \frac{|38{,}042|}{7{,}745} = 4{,}912;\ t_{b_2} = \frac{|1{,}755|}{0{,}595} = 2{,}952;\ t_b = \frac{|-0{,}109|}{9{,}664} = 0{,}011.$$

Для вероятности p=0,95 и k= 22 – 2 – 1 = 19 по таблицам находим:

$$t_{кр}\left(\frac{0{,}05}{2};\ 19\right) = 2{,}093.$$

$$t_{b_0} = 4{,}912 > t_{кр} = 2{,}093;\ t_{b_2} = 2{,}952 > t_{кр} = 2{,}093;\ t_b = 0{,}011 < t_{кр} = 2{,}093.$$

Т.е., с вероятностью 0,95 оценки коэффициентов уравнения регрессии b_0 и b_2 являются статистически значимыми. Это значит, что влияние фактора X_2, т.е., величины расходов на промышленные товары, на уровень доходов действительно существенно.

С той же вероятностью оценка коэффициента уравнения b не является статистически значимой, т.е., влияние фактора D не существенно – уровень доходов несущественно зависит от наличия или отсутствия детей.

Согласно построенному уравнению регрессии наличие детей в семье приводит к уменьшению доходов на 1,109 ед.