Практичне завдання №4

Розв‘язок задач множинної лінійної регресії за допомогою надбудови “Анализ данных”

1. *Відомості з теорії*

Множинною регресією називається модель типу

$$y=f(\mathbf{X},\boldsymbol{\theta})+\varepsilon, \qquad (1)$$

де f - тренд (закономірна складова рівняння регресії, $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots\theta_K)$ – вектор параметрів тренду), ε – похибка регресії (випадкова величина), $\mathbf{X} = (x_1,x_2,\ldots,x_m)$ – вектор факторів (регресорів) і $m>1$.

Лінійною множинною регресією називається модель (1), в який тренд f є лінійною функцією від компонент вектору факторів $\mathbf{X}$:

$$y=a_0+a_1x_1+\ldots+ a_mx_m+\varepsilon. \qquad (2)$$

Вхідними даними для аналізу є просторова або часова вибірка $(\mathbf{X}^{(i)}, y^{(i)})$, $i=\overline{1,n}$. Вибіркові дані (дані спостережень) представляються у вигляді вектору (матриці розмірністю $n\times1$) вибіркових значень залежної змінної (показника) y

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} \qquad (3)$$

і матриці (розмірністю $n\times(m+1)$) значень незалежних змінних (факторів)

$$X=\begin{pmatrix} 1 & x_{11} & x_{12} \ldots & x_{1m} \\ 1 & x_{21} & x_{22} \ldots & x_{2m} \\ \ldots\ldots & \ldots & .. & \ldots \\ 1 & x_{j1} & x_{j2} \ldots & x_{jm} \\ \ldots\ldots & \ldots & .. & \ldots \\ 1 & x_{n1} & x_{n2} \ldots & x_{nm} \end{pmatrix}. \qquad (4)$$

Кожний рядок матриць Y та X містить дані відповідного вибіркового значення; в матриці X введено стовпець, що відповідає фіктивному «нульовому» фактору x_0, значення якого завжди дорівнюють одиниці. Цей фактор введено для зручності подальшого представлення.

Вважається, що модель (2) задовольняється для кожного вибіркового спостереження j

$$y_j=a_0 + \sum_{i=1}^{m} a_i x_{ji} + \varepsilon_j,\ j=1,2,\ldots,n, \qquad (5)$$

тому n рівнянь (5) називають *вибірковими рівняннями*. Ці рівняння можна записати у вигляді єдиного матричного рівняння

$$Y=X\alpha+\varepsilon, \tag{6}$$

де

$$\alpha=\begin{pmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_m \end{pmatrix} \tag{7}$$

– вектор-стовпець коефіцієнтів регресії,

$$\varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix} \tag{8}$$

– вектор випадкових похибок.

Для отримання оцінок

$$\hat{\alpha}=\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{a}_m \end{pmatrix} \tag{9}$$

коефіцієнтів регресії робляться припущення (умови Гауса-Маркова, або *умови «адекватності» регресійної моделі*) щодо компонент вектору похибок регресії, аналогічні відповідним умовам для парної регресії.

I. Випадковість рівнів похибки ε_j, $j=1,2,\ldots,n$.
II. $M[\varepsilon_j]=0$, $j=1,2,\ldots,n$.
III. $\sigma[\varepsilon_j]=\sqrt{D[\varepsilon_j]}=\sigma=const$ (гомоскедастичність рівнів похибки ε_j, $j=1,2,\ldots,n$).
IV. $\varepsilon_j \in N(0,\sigma)$, $j=1,2,\ldots,n$.
V. Некорельованість рівнів похибки ε_i та ε_j: $cov(\varepsilon_i,\varepsilon_j)=0$, $i\neq j$, $i,j=1,2,\ldots,n$.

Згідно з методом найменших квадратів (МНК), оцінки (9) знаходяться як розв'язок оптимізаційної задачі мінімізації квадратичної (евклідової) норми вектору $\boldsymbol{e}$ *залишків регресії*

$$\boldsymbol{e}=\begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{pmatrix}=\begin{pmatrix} y_1-\hat{a}_0-\hat{a}_1x_{11}-\ldots-\hat{a}_mx_{1m} \\ y_2-\hat{a}_0-\hat{a}_1x_{21}-\ldots-\hat{a}_mx_{2m} \\ . \\ . \\ . \\ y_n-\hat{a}_0-\hat{a}_1x_{n1}-\ldots-\hat{a}_mx_{nm} \end{pmatrix}, \tag{10}$$

тобто

$$\|\boldsymbol{e}\|_2^2=\sum_{j=1}^{n}\left[y_j-\left(\hat{a}_0+\sum_{i=1}^{m}\hat{a}_ix_{ji}\right)\right]^2\to min. \tag{11}$$

Для знаходження мінімуму квадратичної форми (11) необхідно прирівняти нулю похідні по параметрах $\hat{a}_i$, $i=0,1,\ldots,m$. В результаті отримаємо систему рівнянь

$$(X^TX)\hat{\alpha}=X^TY, \tag{12}$$

яка називається *системою нормальних рівнянь методу найменших квадратів*. Якщо матриця цієї системи

$$A=(X^TX) \tag{13}$$

є невиродженою, то система (12) має такий розв'язок

$$\hat{\alpha} = (X^TX)^{-1}X^TY = \alpha + (X^TX)^{-1}X^T\varepsilon. \tag{14}$$

Згідно з теоремою Гауса-Маркова, умови I-IV є достатніми для того, щоб оцінки (14) були незміщеними та консистентними оцінками параметрів регресії α. Незміщеною оцінкою дисперсії σ випадкової компоненти регресії буде величина

$$\hat{\sigma}^2 = \frac{1}{n-m-1}\sum_{j=1}^{n}(y_j-\hat{y}_j)^2, \tag{15}$$

де оцінки

$$\hat{y}_j=\hat{a}_0+\sum_{i=1}^{m}\hat{a}_ix_{ji},\ j=1,2,\ldots,n, \tag{16}$$

називають *вирівняними (за регресією) значеннями вибіркових величин* y_j.

Оцінками квадратних коренів з дисперсій величин $\hat{a}_i$ будуть

$$\hat{\sigma}_{\hat{a}_i} = \hat{\sigma}\sqrt{c_{ii}},\ i=1,\ldots,m, \tag{17}$$

де c_{ii} – діагональні елементи матриці C:

$$C=(X^TX)^{-1}. \quad (18)$$

Оцінка значущості моделі. Метод виключення незначущих компонент. Оцінка якості (значущості) лінійної регресійної моделі (2) складається з перевірки значущості моделі в цілому та перевірки значущості кожного з факторів.

Перевірка значущості моделі в цілому полягає у перевірці основної гіпотези H_0: $a_1=a_2=\ldots=a_m=0$ при альтернативі H_1: $a_i\neq 0$ для хоча б одного $i=1,\ldots,m$. F-статистика для критерію перевірки має вигляд

$$F_{\text{розр}}=\frac{\frac{\hat{S}^2}{m}}{\frac{S_R^2}{n-m-1}}, \quad (19)$$

де S_R^2 та $\hat{S}^2$ – залишкова та обумовлена регресією варіація показника відповідно:

$$S_R^2=\sum_{j=1}^{n}\left(y_j-\hat{y}_j\right)^2, \quad (20)$$

$$\hat{S}^2=\sum_{j=1}^{n}\left(\hat{y}_j-\bar{y}\right)^2, \quad (21)$$

$$\bar{y}=\frac{1}{n}\sum_{j=1}^{n}y_j \quad (22)$$

Гіпотезу H_0 слід не відхиляти, якщо $F_{\text{розр}}<F_{(\alpha;m,n-m-1)}$, де $F_{(\alpha;m,n-m-1)}$ – правостороння критична границя розподілу Фішера, і відхиляти, якщо $F_{\text{розр}}\geq F_{(\alpha;m,n-m-1)}$ (α – рівень значущості критерію).

Перевірка значущості i-го фактору полягає у перевірці гіпотези H_0: $a_i=0$ при альтернативі H_1: $a_i\neq 0$. Критеріальна статистика – t-значущість оцінки, тобто

$$t_{i\ \text{розр}}=\frac{\hat{a}_i}{\hat{\sigma}_{\hat{a}_i}}=\frac{\hat{a}_i}{\hat{\sigma}\sqrt{c_{ii}}}. \quad (23)$$

Оцінка є незначущою, якщо її розрахункова t-значущість по модулю не перевищує табличної значущості, тобто двобічної α-критичної границі розподілу Стьюдента

$$|t_{i\ \text{розр}}|\leq t_{\alpha,n-m-1}\ \Rightarrow H_0. \quad (24)$$

Оцінка є значущою в іншому випадку

$$|t_{i\ \text{розр}}|>t_{\alpha,n-m-1}\ \Rightarrow H_1. \quad (25)$$

Вказаний критерій має ймовірність помилки першого роду, що не перевищує α.

Якщо фактор є незначущим, його не варто включати в модель. Послідовна процедура виключення незначущих факторів з рівняння регресії виглядає таким чином.

Знайдемо оцінку з мінімальною значущістю

$$| t_{imin\ \text{розр}} | = \min_{1\le i\le m} |t_{i\ \text{розр}}| . \tag{26}$$

Якщо

$$| t_{imin\ \text{розр}} | \le t_{\alpha,n-m-1}, \tag{27}$$

то виключаємо змінну x_{imin} з рівняння регресії, після чого перераховуємо отриману модель. Процес повторюємо, доки у рівнянні не залишаться лише змінні із значущими оцінками коефіцієнтів регресії.

Прогнозування за рівнянням регресії. Точковий прогноз за рівнянням регресії здійснюється шляхом підставлення значень незалежних змінних $\mathbf{X^{(p)}} = (x_1^{(p)}, x_2^{(p)},\ldots, x_m^{(p)})$ в оцінку детермінованої складової регресійної моделі

$$\hat{y}^{(p)} = \hat{a}_0 + \sum_{i=1}^{m} \hat{a}_i\, x_i^{(p)}. \tag{28}$$

Зрозуміло, що точність прогнозу визначається його дисперсією: чим меншою є дисперсія, тим точніший прогноз. Для визначення розміру довірчого інтервалу прогнозу $[\hat{y}^{(p)}-\Delta y^{(p)};\hat{y}^{(p)}+\Delta y^{(p)}]$ з центром у $\hat{y}^{(p)}$ використаємо оцінку

$$\Delta y^{(p)} = t_{\alpha,n-m-1}\hat{\sigma}_{y^{(p)}}, \tag{29}$$

де

$$\hat{\sigma}_{y^{(p)}} = \hat{\sigma}\sqrt{1+\frac{1}{n}+\sum_{i,k=1}^{m}(x_i^{(p)}-\overline{x}_i)c_{ik}(x_k^{(p)}-\overline{x}_k)}, \tag{30}$$

$$\overline{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji},\ i=1,\ldots,m, \tag{31}$$

c_{ik} – елементи матриці $C=(X^TX)^{-1}$. Ймовірність того, що «точне» значення показника $y^{(p)}$ попаде до інтервалу $[\hat{y}^{(p)}-\Delta y^{(p)};\hat{y}^{(p)}+\Delta y^{(p)}]$, дорівнює $1-\alpha$.

2. Приклад розв'язання задачі

Яка модель найкраще відображає залежність показника y від факторів x_1, x_2, x_3, x_4? Зробити прогноз значення y для даних величин факторів (наведені в останньому рядку таблиці) і знайти його довірчій (α=5%) інтервал.

Таблиця 1. Вхідні дані для розрахунку моделі багатофакторної регресії.

Y	X1	X2	X3	X4
120	1,1	10	53	27,5
150	3,4	12	45	24,7
170	3,5	15	42	23,9
230	4,3	20	34	21,7
255	4,1	28	25	28,4
212	1,8	17	29	15,7
196	2,2	14	48	24,6
178	2,3	12	32	22,3
220	3	19	44	29,7
156	2,4	13	87	25
163	1,7	14	75	19,2
188	1,9	13	43	22
237	1,5	21	29	17,2
135	1,3	11	56	18,3
164	1,8	18	68	12,5
?	3,7	20	70	13

Вхідні дані для розрахунків розмістимо на листі електронної таблиці, як показано на рис. 1.

	A	B	C	D	E
1	Y	X1	X2	X3	X4
2	120	1,1	10	53	27,5
3	150	3,4	12	45	24,7
4	170	3,5	15	42	23,9
5	230	4,3	20	34	21,7
6	255	4,1	28	25	28,4
7	212	1,8	17	29	15,7
8	196	2,2	14	48	24,6
9	178	2,3	12	32	22,3
10	220	3	19	44	29,7
11	156	2,4	13	87	25
12	163	1,7	14	75	19,2
13	188	1,9	13	43	22
14	237	1,5	21	29	17,2
15	135	1,3	11	56	18,3
16	164	1,8	18	68	12,5
17	?	3,7	20	70	13

Рис. 1. Розміщення вхідних даних для розрахунку регресійної моделі.

Спочатку побудуємо регресійну модель на основі всіх чотирьох факторів x_1–x_4. Використовуємо надбудову Аналіз даних->Регресія (Data Analysis->Regression). У діалоговому вікні «Регресія» (рис. 2) введемо вхідні дані для розрахунку.

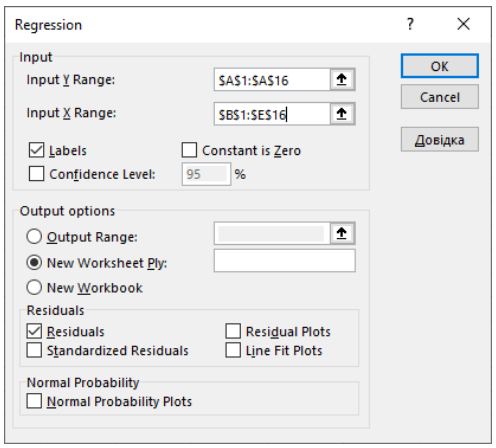

Рис. 2. Вигляд діалогового вікна «Регресія».

В якості вхідного діапазону даних для y позначимо $A$1:$A$16, діапазону x – $B$1:$E$16. Імена заголовків для факторів і показнику включаємо у відповідні діапазони (галочка «Мітки» (“Labels”)). Результати розрахунків розмістимо на новому робочому листі книги (рис. 3).

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0,914356344							
5	R Square	0,836047524							
6	Adjusted R Square	0,770466534							
7	Standard Error	18,86332329							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	18144,68368	4536,171	12,74832	0,000613677			
13	Residual	10	3558,249654	355,825					
14	Total	14	21702,93333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
17	Intercept	120,9279837	39,2158404	3,083651	0,011568	33,5496461	208,306321	33,5496461	208,306321
18	X1	-0,111115701	6,907406203	-0,01609	0,987482	-15,50177583	15,2795444	-15,5017758	15,2795444
19	X2	6,006820167	1,377946827	4,359254	0,001423	2,936563306	9,07707703	2,936563306	9,07707703
20	X3	-0,679564744	0,321564923	-2,1133	0,060704	-1,396056042	0,03692655	-1,39605604	0,03692655
21	X4	0,069096281	1,214039372	0,056914	0,955735	-2,635952011	2,77414457	-2,63595201	2,77414457
22									
23									
24									
25	RESIDUAL OUTPUT								
26									
27	*Observation*	*Predicted Y*	*Residuals*						
28	1	146,7571744	-26,75717442						
29	2	163,758297	-13,75829701						

Рис 3. Результати розрахунку регресійної моделі із включенням усіх чотирьох факторів x_1–x_4.

Проаналізуємо результати розрахунку регресійної моделі. Множинний (вибірковий) коефіцієнт кореляції $\hat{\rho}_{y\bullet(x_1,\ldots,x_4)}$, показаний у комірці B4, дорівнює 0,914356344, що характеризує достатньо високий рівень наближення показника y лінійними комбінаціями факторів x_1–x_4. Для перевірки значущості моделі в цілому розглянемо значення статистики $F_{\text{розр}}$ критерію Фішера, обчислене за формулою (19). Це значення розташоване в комірці E12. Для перевірки критерію розрахуємо теоретичне значення F-критерію з

рівнем значущості α=5%, застосувавши формулу =F.INV.RT(0,05;4;10) (російськомовний варіант функції F.ОБР.ПХ, рис. 4).

MS	*F*	*Significance F*
4536,17092	12,74832	0,000613677
355,824965		
Fα,m,n-m-1	3,47805	

Рис. 4. Перевірка значущості регресійної моделі в цілому.

Оскільки $F_{\text{розр}} = 12{,}74832 > F_{(\alpha;m,n-m-1)} = F_{(0{,}05;4,15-4-1)} = F_{(0{,}05;4,10)} = 3{,}47805$, то гіпотезу H_0: $a_1=a_2=\ldots=a_m=0$ ми відхиляємо і робимо висновок, що модель в цілому є значущою. Аналогічний результат можемо здобути, використавши так зване *p*-значення критерію Фішера (ймовірність не відхиляти нульову гіпотезу за критерієм на основі даних спостережень, якщо вона справджується). Це значення, розташоване в комірці F12 (*Significance F* або *«Значущість F»*), дорівнює 0,000613677, і, оскільки воно менше заданого рівня значущості α=0,05, гіпотезу H_0 відхиляємо.

Переходимо до перевірки значущості кожного з факторів a_i, $i=1,2,\ldots,m$. Вибіркові значення коефіцієнтів $\hat{a}_i$, $i=0,1,\ldots,m$, розташовані в комірках B17-B21. Відповідні розрахункові значення *t*-статистики знаходяться в комірках D17-D21. Теоретичне значення *t*-статистики $t_{\alpha,n-m-1} = t_{0{,}05,15-4-1} = t_{0{,}05,10}$ розрахуємо за формулою =T.INV.2T(0,05;10) (російськомовний варіант СТЬЮДЕНТ.ОБР.2Х, рис. 5).

Standard Error	*t Stat*	*P-value*
39,2158404	3,08365147	0,011568
6,907406203	-0,0160865	0,987482
1,377946827	4,35925396	0,001423
0,321564923	-2,113305	0,060704
1,214039372	0,05691437	0,955735
tα,n-m-1	2,22813885	

Рис. 5. Перевірка значущості кожного з факторів моделі.

Нерівність (24) $|t_{i\ \text{розр}}| \le t_{\alpha,n-m-1}$ виконується для $i=1,3,4$, тобто ці фактори нам варто вважати незначущими. Аналогічний результат можна отримати на основі *p*-значень, розташованих в комірках E17-E21. *p*-значення для вказаних факторів більші за рівень значущості α=0,05, тому гіпотези H_0: $a_i=0$ для $i=1,3,4$ не відхиляємо для даного рівня значущості.

З факторів x_1, x_3, x_4 вибираємо фактор з мінімальним значенням $|t_{i\ \text{розр}}|$ модуля розрахункового значення *t*-критерію (або, що те ж саме, фактор з максимальною величиною *p*-значення). Це буде фактор x_1 з відповідним значенням $|t_{1\ \text{розр}}| = 0{,}0160865$ (відповідна величина *p*-значення дорівнює 0,987482).

Виключаємо фактор x_1 з моделі і робимо її перерахунок із змінними x_2, x_3, x_4 в якості регресорів. Результати розрахунку оновленої моделі представлено на рис. 6.

D21 =T.INV.2T(0,05;11)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0,914354024							
5	R Square	0,836043281							
6	Adjusted R Sc	0,791327813							
7	Standard Erro	17,98570577							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	18144,5916	6048,1972	18,69696	0,000125636			
13	Residual	11	3558,341733	323,485612					
14	Total	14	21702,93333						
15				Fα,m,n-m-1	3,587434				
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
17	Intercept	121,0269654	36,92818877	3,27735991	0,007368	39,74856993	202,305361	39,7485699	202,3053609
18	X2	5,995863239	1,142103303	5,24984318	0,000273	3,482110817	8,50961566	3,48211082	8,509615661
19	X3	-0,678993678	0,304729999	-2,2281813	0,047678	-1,349699883	-0,0082875	-1,3496999	-0,00828747
20	X4	0,059096613	0,994305594	0,05943506	0,953672	-2,129355244	2,24754847	-2,1293552	2,24754847
21			tα,n-m-1	2,20098516					
22									
23									
24	RESIDUAL OUTPUT								
25									
26	*Observation*	*Predicted Y*	*Residuals*						
27	1	146,6240897	-26,6240897						
28	2	163,8822951	-13,8822951						

Рис. 6. Результати розрахунку на другому етапі процесу виключення незначущих факторів. В модель включені змінні x_2, x_3, x_4.

Аналізуючи результати розрахунків, бачимо, що множинний коефіцієнт кореляції залишився на приблизно такому ж самому рівні ($\hat{\rho}_{y\bullet(x_1,\ldots,x_4)}$ =0,914354024). F-статистика для оновленої моделі збільшилася і складає $F_{\text{розр}}$=18,69695892. Теоретичне значення F-критерію $F_{(\alpha;m,n-m-1)} = F_{(0,05;3,15-3-1)} = F_{(0,05;3,11)}$ = 3,587433702 (не варто забувати, що кількість факторів зменшилася і вже складає m=3); $F_{\text{розр}}>F_{(0,05;3,11)}$, отже, нульову гіпотезу відхиляємо і модель у цілому вважається значущою.

Перевіряючи значущість кожного з факторів, отримаємо, що гіпотезу H_0: a_i=0 слід не відхиляти лише для фактору x_4: $|\, t_{4\ \text{розр}}\,|$=0,05943506<$t_{\alpha,n-m-1}$= $t_{0,05,15-3-1} = t_{0,05,11}$ = 2,20098516 (відповідне p-значення дорівнює 0,953672>0,05). Тобто, цей фактор вважається незначущим і підлягає виключенню з моделі.

Виключаємо змінну x_4 і робимо перерахунок з урахуванням факторів x_2, x_3 (рис. 7).

D20 =T.INV.2T(0,05;12)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0,914325231							
5	R Square	0,835990629							
6	Adjusted R S(	0,808655733							
7	Standard Errc	17,22276704							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	18143,44888	9071,72444	30,5832754	1,94631E-05			
13	Residual	12	3559,484454	296,6237045					
14	Total	14	21702,93333						
15				Fα,m,n-m-1	3,88529383				
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
17	Intercept	122,4428076	27,02040125	4,531494795	0,00068792	63,57041071	181,3152045	63,5704107	181,315205
18	X2	5,994901337	1,093546419	5,482073036	0,00014024	3,612268369	8,377534304	3,61226837	8,3775343
19	X3	-0,680892579	0,290195377	-2,3463247	0,03695864	-1,313173989	-0,04861117	-1,31317399	-0,04861117
20			tα,n-m-1	2,17881283					
21									
22									
23	RESIDUAL OUTPUT								
24									
25	*Observation*	*Predicted Y*	*Residuals*						
26	1	146,3045143	-26,3045143						
27	2	163,7414576	-13,7414576						

Рис. 7. Розрахунок моделі з урахуванням змінних x_2, x_3 у якості факторів.

Аналізуючи результати, дійдемо висновку, що тепер всі фактори в моделі є значущими, і найкращою для аналізу та прогнозування буде модель $y=a_0+a_2x_2+a_3x_3+\varepsilon$, де коефіцієнти моделі мають такі оцінки: $\hat{a}_0$=122,4428076, $\hat{a}_2$=5,994901337, $\hat{a}_3$= –0,680892579.

Виконаємо прогнозування на базі побудованої моделі значення показника y для заданого вектору факторів $\mathbf{X_p}$={3.7, 20, 70, 13}. Приклад розміщення вхідних даних і результатів обчислень на листі електронної таблиці наведено на рис. 8.

H26 =G19*H19*SQRT(1+1/L2+SUMPRODUCT(I22:I24;MMULT(D22:F24;I22:I24)))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Y	X0	X2	X3	Y^	ei						n	m
2	120	1	10	53	146,3045	-26,3045						15	2
3	150	1	12	45	163,7415	-13,7415							
4	170	1	15	42	183,7688	-13,7688							
5	230	1	20	34	219,1905	10,80951							
6	255	1	28	25	273,2777	-18,2777							
7	212	1	17	29	204,6102	7,389754							
8	196	1	14	48	173,6886	22,31142							
9	178	1	12	32	172,5931	5,406939							
10	220	1	19	44	206,3867	13,61334							
11	156	1	13	87	141,1389	14,86113							
12	163	1	14	75	155,3045	7,695517							
13	188	1	13	43	171,0981	16,90186							
14	237	1	21	29	228,5899	8,410149							
15	135	1	11	56	150,2567	-15,2567							
16	164	1	18	68	184,0503	-20,0503							
17	Xcep	1	15,8	47,33333									
18			a3	a2	a0		σ^	tα,n-m-1					
19			-0,68089	5,994901	122,4428		15,306827	2,1788128					
20													
21	A			$C=A^{-1}$									
22	15	237	710	2,461375	-0,0875	-0,02138		X0-X0cep	0				
23	237	4063	10654	-0,0875	0,004032	0,000503		X2p-X2cep	4,2				
24	710	10654	38128	-0,02138	0,000503	0,000284		X3p-X3cep	22,66667				
25		X0p	X1p	X2p	X3p	X4p	Yp^	ΔYp^	Yp^-	Yp^+			
26		1	3,7	20	70	13	194,67835	39,16969	155,5087	233,848			

Рис. 8. Розташування вхідних даних та результатів обчислень на листі електронної таблиці для прогнозування значень показника y.

З набору вхідних даних залишимо стовпці із значеннями y, x_0 (фіктивний фактор), x_2, x_3. Обчислимо коефіцієнти регресії a_0, a_2, a_3 в комірках C19:E19, виділивши цей діапазон і увівши формулу =LINEST(A2:A16;C2:D16) (російськомовна назва функції ЛИНЕЙН) і натиснувши комбінацію клавіш Ctrl-Shift-Enter. Вирівняні за регресією вибіркові значення показника $\hat{y}_j$ розташуємо в комірках E2:E16, увівши в комірці E2 формулу =$E$19+$D$19*C2+$C$19*D2 і скопіювавши її на весь діапазон. Залишки регресії e_j розмістимо в стовпці F за допомогою формули в комірці F2: =A2-E2, скопійованої на весь діапазон. Розмір вибірки n та кількість регресорів m розмістимо у комірках L2 та M2 відповідно.

Обчислимо значення вибіркового середньоквадратичного відхилення $\hat{\sigma}$ похибки регресії за формулою (15), увівши в комірку G19 формулу =SQRT(SUMSQ(F2:F16)/L2-M2-1) (російськомовні назви функцій КОРЕНЬ та СУММКВ). У сусідню комірку H19 введемо формулу для α-границі $t_{\alpha,n-m-1}$ розподілу Стьюдента =T.INV.2T(0,05;L2-M2-1) (російськомовний варіант СТЬЮДЕНТ.ОБР.2Х).

Тепер залишилося для реалізації формули (30) для $\hat{\sigma}_{y^{(p)}}$ обчислити елементі матриці $C=A^{-1}=(X^TX)^{-1}$ (формула (18)). Матрицю A розташуємо в комірках A22:C24, виділивши їх і набравши формулу =MMULT(TRANSPOSE(B2:D16);B2:D16) (МУМНОЖ та ТРАНСП в російськомовному варіанті). Пам’ятаємо, що всі матричні операції завершуються натисканням комбінації клавіш Ctrl-Shift-Enter. Обернену до матриці A матрицю C розмістимо в діапазоні D22:F24, увівши формулу =MINVERSE(A22:C24) (МОБР). Середні значення факторів розмістимо в комірках B17:D17, увівши в комірку B17 формулу =AVERAGE(B2:B16) і «протягнувши» її на весь зазначений діапазон. Вектор відхилень компонент вектору прогнозного значення факторів від їх вибіркових середніх $x_i^{(p)}-\overline{x}_i$, що використовується для обчислення суми під знаком кореня у формулі (29), розташуємо в діапазоні I22:I24. Для цього в комірку I22 введемо формулу =B26-B17, у комірку I23 – формулу =D26-C17, у комірку I24 – формулу =E26-D17.

Все готово для обчислення точкового значення прогнозу $\hat{y}^{(p)}=\hat{a}_0+\hat{a}_2x_2^{(p)}+\hat{a}_3x_3^{(p)}$ та його довірчого інтервалу $\Delta y^{(p)}$ за формулою (29). Значення факторів, для яких необхідно знайти прогноз, розмістимо в діапазоні B26:F26. В комірку G26 введемо формулу для точкового прогнозу =$E$19+$D$19*D26+$C$19*E26, а в комірку H26 – формулу для $\Delta y^{(p)}$: =G19*H19*SQRT(1+1/L2+SUMPRODUCT(I22:I24;MMULT(D22:F24;I22:I24))) (КОРЕНЬ, СУММПРОИЗВ, МУМНОЖ). Нижню $\hat{y}^{(p)}-\Delta y^{(p)}$ та верхню $\hat{y}^{(p)}+\Delta y^{(p)}$ границі прогнозу обчислимо в комірках I26 та J26 шляхом введення формул =G26-H26 та =G26+H26 відповідно.

Отже, прогнозне значення показнику $\hat{y}^{(p)}$ дорівнюватиме 194,678353; з ймовірністю 95% «точне» значення $y^{(p)}$ попаде у інтервал [155,509;233,848].

3. Задачі

Таблиця відповідності номерів варіантів порядковим номерам студентів за журналом групи

Номер студента в журналі групи	Номер варіанта лабораторної роботи	Номер студента в журналі групи	Номер варіанта лабораторної роботи	Номер студента в журналі групи	Номер варіанта лабораторної роботи
1	1	10	4	19	1
2	2	11	5	20	2
3	3	12	6	21	3
4	4	13	1	22	4
5	5	14	2	23	5
6	6	15	3	24	6
7	1	16	4	25	1
8	2	17	5	26	2
9	3	18	6	27	3

Варіант 1.

Знайдіть найкращу модель для прогнозування значень X, якщо відомі величини Y і Z. Чи виправдане одночасна присутність у рівнянні Y і Z?

Джерело: Дрейпер Н., Сміт Г., Прикладний регресійний аналіз: у 2-х кн. Кн.1, 2-ге вид., перероб. і доп.- М.: Фінанси і статистика, 1986,с. 262

№	*X*	*Y*	*Z*
1	*1,52*	*98*	*77*
2	*1,41*	*76*	*139*
3	*1,16*	*58*	*179*
4	*1.45*	*94*	*95*
5	*1,24*	*73*	*142*
6	*1,21*	*57*	*186*
7	*1,63*	*97*	*82*
8	*1,38*	*91*	*100*
9	*1,37*	*79*	*125*

10	*1,36*	*92*	*96*
11	*1,4*	*92*	*99*
12	*1,03*	*54*	*190*

Варіант 2.

Оцінювач має наступні дані про характеристики одинадцяти будинків (в одному районі міста), орендованих або куплених фірмами.

Необхідно знайти найкращу модель для оцінки 12-го будинку.

№	**Загальна площа, (кв. м.)**	**Кількість офісів**	**Кількість входів**	**Термін експлуатації (рік)**	**Вартість (у.г.о.)**
	X1	**X2**	**X3**	**X4**	**Y**
1	2310	2	2	20	142000
2	2333	2	2	12	144000
3	2356	3	1.5	33	151000
4	2379	3	2	43	150000
5	2402	2	3	53	139000
6	2425	4	2	23	169000
7	2448	2	1,5	99	126000
8	2471	2	2	34	142900
9	2494	3	3	23	163000
10	2517	4	4	55	169000
11	2540	2	3	22	149000
12	2500	3	2	25	?

(Джерело: Допомога до функції ЛИНЕЙН пакета EXCEL Microsoft Office 97)

Варіант 3.

Знайдіть найкращу модель для прогнозування значень змінної X5, якщо

відомі X1, X2, X3, X4. (Джерело: Дрейпер Н., Сміт Г. Кн. 2, стор. 283-231). Розв‘яжіть задачу методом виключення і кроковим методом. Порівняйте результати і зробіть висновки.

№	*X1*	*X2*	*X3*	*X4*	*X5*
1	*7*	*26*	*6*	*60*	*78,5*
2	*1*	*29*	*15*	*52*	*74,3*
3	*11*	*56*	*8*	*20*	*104,3*
4	*11*	*31*	*8*	*47*	*87,6*
5	*7*	*52*	*6*	*33*	*95,9*
6	*11*	*55*	*9*	*22*	*109,2*
7	*3*	*71*	*17*	*6*	*102,7*
8	*1*	*31*	*21*	*44*	*72,5*
9	*2*	*54*	*18*	*22*	*93,1*
10	*21*	*47*	*4*	*26*	*115,9*
11	*1*	*40*	*23*	*34*	*83,8*
12	*11*	*66*	*9*	*12*	*113,3*
13	*10*	*68*	*8*	*12*	*109,4*

Варіант 4.

Маємо дані стоматологічного обстеження робітників-чоловіків.

ALTER	S	PU	ZB	CPITN
20	3	3	3	2,17
61	2	2	4	3,6
27	3	3	3	2,33
16	2	2	4	2
54	2	2	4	3,6
37	2	2	4	3,33
58	2	2	4	4
17	2	2	4	3
56	2	3	3	2
31	2	2	4	2,83
20	3	3	3	1,33
50	3	3	3	2,33
37	2	2	4	3,83

27	3	3	3	2
51	3	3	3	2,67
47	2	2	4	3
25	2	2	3	2,5
35	2	2	4	2,83
24	2	2	4	2,83
60	2	3	3	2
30	3	3	3	1,67
27	3	3	3	1,5
25	3	3	3	1,33
42	2	3	3	2,5
37	2	2	3	3
22	3	2	4	3,17
35	3	2	3	3
35	3	2	3	2,6
32	2	2	3	2,83
26	2	2	3	2,5

Загальний рівень стану зубів охарактеризований параметром CPITN, який може приймати значення від 0 до 4, де 0 відповідає здоровому стану, 4 – найвищому ступеню розвитку захворювання. Інші параметри, які приймаються за незалежні змінні, розшифровуються таким чином

ALTER	Вік особи
S	Освіта (1 – спеціальна шкільна, 2 – неповна шкільна, 3 – середня, 4 – атестат зрілості, 5 – вища)
PU	Періодичність чистки зубів (1 – менше одного разу на день, 2 – один раз на день, 3 – два рази на день, 4 – більше двох разів на день)
ZB	Зміна зубної щітки (1 – кожний місяць, 2 – кожні три місяці, 3 – раз у півроку, 4 – ще рідше)

Побудуйте найкращу модель зв‘язку між параметром CPITN та

незалежними змінними. Проаналізуйте модель, оцініть її якість.

(Адаптовано по джерелу «SPSS: искусство обработки информации»/А. Бююль, П. Цефель, ДиаСофт, 2002).

Варіант 5.

Результати обстеження десяти статистично однорідних філіалів фірми наведені в таблиці.

№ філії	Виробництво праці (y)	Фондооснащеність (x_1)	Енергооснащеність (x_2)
1	74	33	56
2	84	34	58
3	73	36	67
4	93	35	70
5	56	33	73
6	71	37	77
7	117	39	78
8	111	42	99
9	135	43	93
10	125	44	96

Побудувати модель множинної лінійної регресії, проаналізувати її та оцінити якість. Зробити прогноз значення показника y для x_1=40, x_2=85 та обчислити його довірчій інтервал.

Варіант 6.

Двадцять осіб із зайвою вагою (11 чоловіків та 9 жінок) виявили бажання схуднути і для цього взялись дотримуватись певної дієти. Одинадцять випробуваних додатково вступили у товариство, де процес схуднення стимулюється за допомогою спеціальних лекцій та інших мотивуючих методів. Для всіх осіб, що тестуються, були зняті показники зросту та ваги до та після проходження курсу. Ефективність процесу далі обчислювалась за допомогою індексу Брока, в якому фактична вага була віднесена до нормальної ваги (яка, в свою чергу, обчислювалась як різниця зросту, вираженому в см., та числа 100). Результати тестування наведені в таблиці

BEH	G	GR	BROCAB
1	1	175	4,133333
1	2	164	4,375
1	1	183	2,168675
1	1	180	4,5
1	2	159	4,40678
1	1	187	7,356322
1	2	165	10,15385
1	1	177	5,714286
1	2	162	3,387097
2	1	185	13,29412
2	2	156	13,57143
2	2	161	11,63934
2	1	172	8,055556
2	1	180	15,25
2	2	164	9,0625
2	1	170	12
2	2	166	12,12121
2	2	158	9,482759
2	1	190	12,33333
2	1	170	10,71429

Тут змінна beh вказує на групу (1 – дієта, 2 – дієта + товариство бажаючих схуднути), змінна g – на стать (1 – чол., 2 – жін.), gr – зріст, brocab – зменшення індексу Брока, що відображає ефективність процесу схуднення. Чи існує залежність між показником brocab та переліченими факторами? Побудувати лінійну регресійну модель та проаналізувати її.

(Адаптовано по джерелу «SPSS: искусство обработки информации»/А. Бююль, П. Цефель, ДиаСофт, 2002).