Лекція 3

Класична лінійна багатофакторна модель.

Кафедра інформаційних технологій
доцент Бесклінська О.П.

Зміст:

1. Методи побудови багатофакторної регресійної моделі
2. Етапи дослідження загальної лінійної моделі множинної регресії
3. Приклад параметризації та дослідження багатофакторної регресійної моделі

1. Методи побудови багатофакторної регресійної моделі

1. Метод усіх можливих регресій
2. Метод виключень
3. Покроковий регресійний метод

2. Етапи дослідження загальної лінійної моделі множинної регресії

Розглядається багатофакторна лінійна регресійна модель

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m$$

Для дослідження моделі слід виконати такі кроки.

1 За даними спостережень оцінити параметри $a_1, a_2, \ldots, a_m$

2 2. Для перевірки адекватності отриманої моделі обчислити:

а) залишки моделі — розбіжності між спостереженими та розрахунковими значеннями залежної змінної $u_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$

б) відносну похибку залишків та її середнє значення

в) залишкову дисперсію

г) коефіцієнт детермінації

д) вибірковий коефіцієнт множинної кореляції

3 Перевірити статистичну значущість отриманих результатів:

а) перевірити адекватність моделі загалом:
за допомогою *F*-критерію Фішера перевірити гіпотезу

$$H_0: a_1 = a_2 = \ldots = a_m = 0$$

проти альтернативної

H_A: існує хоча б один коефіцієнт $a_j \neq 0$

б) перевірити значущість коефіцієнта множинної кореляції, тобто розглянути гіпотезу

$$H_0 : R = 0$$

в) перевірити істотність коефіцієнтів регресії: за допомогою t-критерію Стьюдента перевірити гіпотезу

H_0 : $a_j = 0$ для всіх $j = 1, 2, \ldots, m$

проти відповідних альтернативних гіпотез

H_A : $a_j \neq 0$ для всіх $j = 1, 2, \ldots, m$;

4 **Обчислити та інтерпретувати коефіцієнти еластичності**

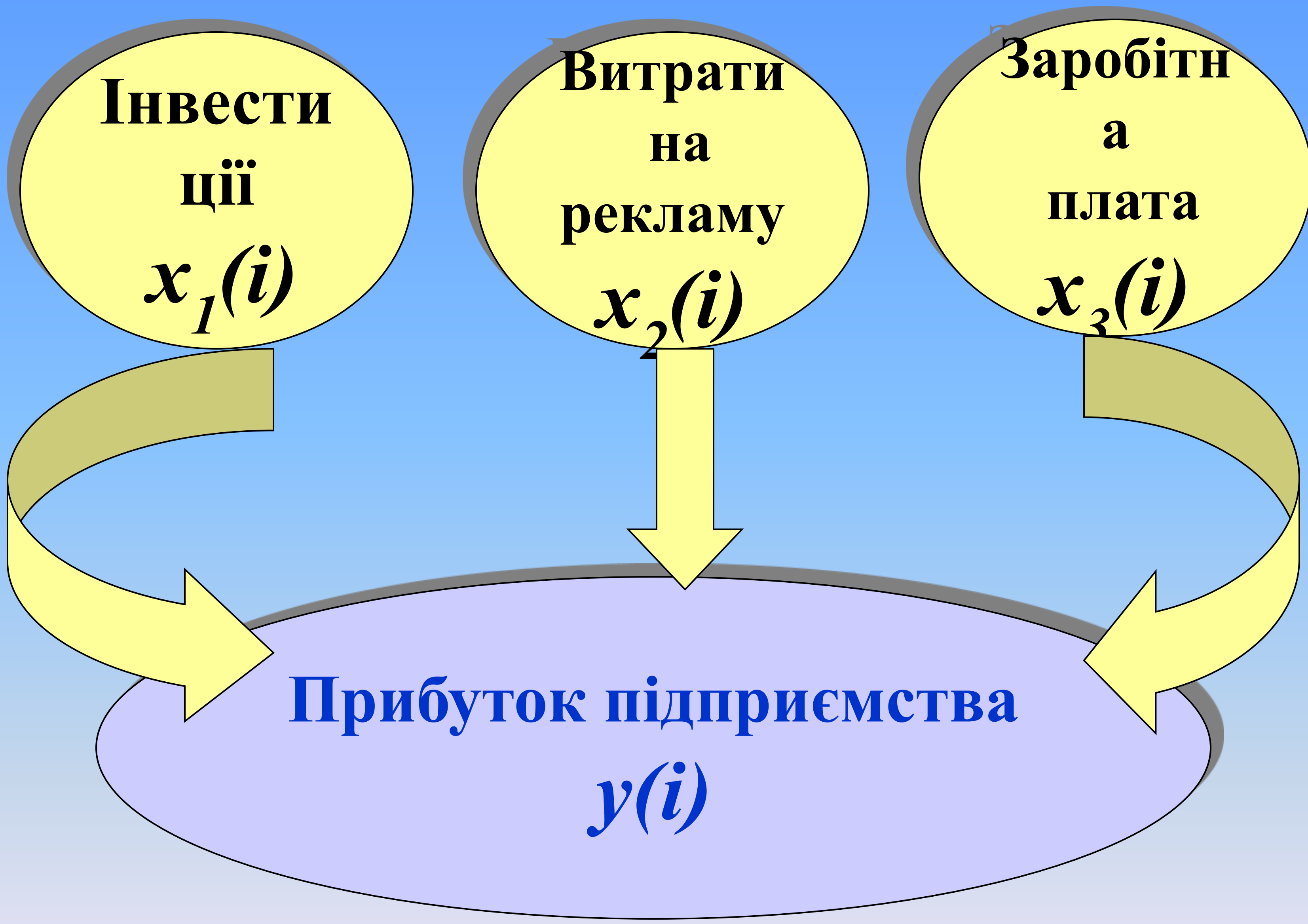
5. **Визначити довірчі інтервали регресії при рівні значущості α.**

6. **Побудувати довірчі інтервали для параметрів регресії.**

7. Обчислити прогнозні значення y_p

за значеннями $x_{1p}, x_{2p}, \ldots, x_{mp}$,
що перебувають за межами
базового періоду,
і знайти межі довірчих інтервалів
індивідуальних прогнозованих значень
і межі довірчих інтервалів
середнього прогнозу.

3. Приклад параметризації та дослідження багатофакторної регресійної моделі


Інвестиції
$x_1(i)$
Витрати на рекламу
$x_2(i)$
Заробітна плата
$x_3(i)$
Прибуток підприємства
$y(i)$

Припустимо, що між економічним показником *y* і факторами x_1, x_2, x_3 існує лінійний зв'язок.

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3$$

$$\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3$$

параметри моделі, які потрібно оцінити

Вихідні дані в умовних одиницях

Номер спост.	$x_1(i)$	$x_2(i)$	$x_3(i)$	$y(i)$
1	17,37	5,28	1,42	15,7
2	18,24	6,47	1,58	17,34
3	22,47	6,98	1,98	21,57
4	18,47	7,05	2,04	33,5
5	16,82	7,94	2,38	32,30
……	……	……	……	……
14	35,67	18,47	8,58	62,22
15	47,87	19,64	9,47	77,58

1

Знайдемо МНК-оцінки параметрів моделі.

$$Y := \begin{pmatrix} 15.7 \\ 17.34 \\ 21.57 \\ 33.5 \\ 32.3 \\ 37.9 \\ 40.78 \\ 48.02 \\ 43.3 \\ 49.57 \\ 52.14 \\ 55.17 \\ 59.18 \\ 62.22 \\ 77.58 \end{pmatrix}
\qquad
\begin{array}{c} \begin{array}{cccc} & x_1(i) & x_2(i) & x_3(i) \end{array} \\
X := \begin{pmatrix} 1 & 17.37 & 5.28 & 1.42 \\ 1 & 18.24 & 6.47 & 1.58 \\ 1 & 22.47 & 6.98 & 1.98 \\ 1 & 18.47 & 7.05 & 2.04 \\ 1 & 16.82 & 7.94 & 2.38 \\ 1 & 17.6 & 8.12 & 3.48 \\ 1 & 17.12 & 8.69 & 3.07 \\ 1 & 19.81 & 9.31 & 3.84 \\ 1 & 18.67 & 10.45 & 4.28 \\ 1 & 20.83 & 10.47 & 4.67 \\ 1 & 22.84 & 13.48 & 5.98 \\ 1 & 28.85 & 15.78 & 6.51 \\ 1 & 29.61 & 17.65 & 7.82 \\ 1 & 35.67 & 18.47 & 8.58 \\ 1 & 47.87 & 19.64 & 9.47 \end{pmatrix} \end{array}$$

Обчислимо оцінки регресійних коефіцієнтів за формулою

$$\hat{a} = (X^T X)^{-1} X^T Y$$

де X^T — транспонована матриця X

Виконавши обчислення, одержимо параметри моделі:

N=15	Regression Summary for Dependent Variable: y (Данные) R= ,95622179 R?= ,91436011 Adjusted R?= ,89100378 F(3,11)=39,148 p<,00000 Std.Error of estimate: 5,7357					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			26,10789	8,407584	3,10528	0,010009
x1	-0,126364	0,188376	-0,25180	0,375368	-0,67081	0,516179
x2	-0,744814	0,679027	-2,72767	2,486744	-1,09688	0,296124
x3	1,797200	0,666449	11,85602	4,396525	2,69668	0,020780

$$\mathbf{a} = \begin{pmatrix} 26.108 \\ -0.252 \\ -2.728 \\ 11.856 \end{pmatrix} \begin{array}{l} a_0 \\ a_1 \\ a_2 \\ a_3 \end{array}$$

Функція регресії з урахуванням знайдених оцінок параметрів моделі набуває вигляду

$$y = 26{,}108 - 0{,}252x_1 - 2{,}728x_2 + 11{,}856x_3$$

2

Для перевірки адекватності отриманої моделі обчислимо:

а) Залишки $u_i = y_i - \hat{y}_i$

б) Відносну похибку розрахункових значень регресії:

$$\delta_i = \frac{u_i}{y_i} \cdot 100\%$$

середнє значення відносної похибки

$$\bar{\delta} = \frac{\sum_{i=1}^{n} \delta_i}{n}$$

У нас

$$\bar{\delta} = -0{,}52783$$

в) Обчислимо середньоквадратичну помилку дисперсії збурень

$$\hat{S}_u = \hat{\sigma}_u = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n-m-1}}$$

У нас $\quad \hat{S}_u = 5{,}7357$

N=15	Regression Summary for Dependent Variable: y (Данные) R= ,95622179 R?= ,91436011 Adjusted R?= [illegible] F(3,11)=39,148 p<,0000[illegible] Std.Error of estimate: 5,7357					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			26,10789	8,407584	3,10528	0,010009
x1	-0,126364	0,188376	-0,25180	0,375368	-0,67081	0,516179
x2	-0,744814	0,679027	-2,72767	2,486744	-1,09688	0,296124
x3	1,797200	0,666449	11,85602	4,396525	2,69668	0,020780

г) Перевіримо тісноту загального зв'язку (впливу) незалежних змінних на залежну змінну. Для цього треба обчислити коефіцієнт детермінації за формулою

$$R^2 = 1 - \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

N=15	Regression Summary for Dependent Variable: y (Данные) R= ,95622179 R?= ,91436011 Adjusted R?= ,89100378 F(3,11)=39,140 p<,00000 Std.Error of estimate: 5,7357					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			26,10789	8,407584	3,10528	0,010009
x1	-0,126364	0,188376	-0,25180	0,375368	-0,67081	0,516179
x2	-0,744814	0,679027	-2,72767	2,486744	-1,09688	0,296124
x3	1,797200	0,666449	11,85602	4,396525	2,69668	0,020780

У нас

$$R^2 = 0{,}91436$$

Висновок: чим ближчий він до одиниці, тим більше варіація залежної змінної *y* визначається варіацією незалежної змінної *x*

(є тісний зв'язок між залежною та незалежними змінними).

д) Перевіримо на значущість вибірковий коефіцієнт кореляції.

Для цього обчислимо $R=\sqrt{R^2}$- коефіцієнт кореляції (характеризує тісноту лінійного зв'язку всіх незалежних факторів x_i із залежною змінною y).

У нас

$$R=\sqrt{R^2}=\sqrt{0{,}91436}=0{,}956222$$

3

Перевіримо статистичну значущість отриманих результатів.

а) Обчислимо *F*-статистику за формулою

$$F_{експ} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$$

Знайти табличне значення: $F(m, n-m-1, \alpha)$

і порівняти його з обчисленою *F* – статистикою: якщо

$$F_{експ} > F(m, n-m-1, \alpha)$$

то гіпотеза відхиляється, інакше приймається.

У нас

Маємо $F_{експ}$**=39,14827,**

табличне значення: $F(3;11;0,05) = 3,59$

Порівняємо його з обчисленою ***F*-статистикою.**

Оскільки $F_{експ} > F(3;11;0,05)$

нульова гіпотеза відхиляється, тобто коефіцієнти регресії є значущими.

Regression Summary for Dependent Variable: y (Данные)
R= ,95622179 R?= ,91436011 Adjusted R?= ,89100378
F(3,11)=39,148 p<,00000 Std.Error of estimate: 5,7357

N=15	Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			26,10789	8,407584	3,10528	0,010009
x1	-0,126364	0,188376	-0,25180	0,375368	-0,67081	0,516179
x2	-0,744814	0,679027	-2,72767	2,486744	-1,09688	0,296124
x3	1,797200	0,666449	11,85602	4,396525	2,69668	0,020780

б) Обчислимо *t*-статистику за формулою

$$t = \frac{R\sqrt{n-m-1}}{\sqrt{1-R^2}}$$

Якщо $|t| > t_{табл}(\alpha/2, n-m-1)$

де $t_{табл}(\alpha/2, n-m-1)$

відповідне табличне значення *t*-розподілу з (*n-m-1*) ступенями свободи, то можна зробити висновок про значущість коефіцієнта кореляції між залежною і незалежними змінними моделі.

У нас

Маємо t = 37,03215.
Відповідне табличне значення

$$t_{табл}(0{,}025;11) = 2{,}593097$$

Оскільки $\left|t\right| > t_{табл}(0{,}025;11)$

можна зробити висновок про достовірність коефіцієнта кореляції, який характеризує тісноту зв'язку між залежною та незалежними змінними моделі.

Для вибраного рівня значущості α і відповідного ступеня вільності $k=n-m-1$ записати межі надійності для множинного коефіцієнта кореляції R:

(R-ΔR; R+ΔR), **де**

$$\Delta R = t_{\alpha/2,k} \cdot \frac{1-R}{\sqrt{n}}$$

У нас

Маємо

$$\Delta R = 2{,}593 \cdot \frac{1 - 0{,}956222}{\sqrt{15}} = 0{,}029311$$

отже

$$(R - \Delta R\,;\, R + \Delta R) = (0{,}926911\,;\, 0{,}985533)$$

в) Перевіримо значущість окремих коефіцієнтів регресії.

Визначимо *t*-статистику за формулою

$$t_j = \frac{a_j}{\sqrt{\sigma_u^2 c_{jj}}} = \frac{a_j}{S_{a_j}}$$

де c_{jj} - діагональний елемент матриці $(X^T X)^{-1}$,

S_{a_j} - стандартизована помилка оцінки параметра моделі.

Значення t_j-критерію порівнюється з табличними при $k = n\text{-}m\text{-}1$ ступенях свободи і рівні значущості α:

якщо $|t_j| > t_{\alpha/2,k}$, то відповідна оцінка параметра регресійної моделі є значуща; інакше приймаємо гіпотезу про рівність a_j нулю

	Regression Summary for Dependent Variable: y (Данные) R= ,95622179 R?= ,91436011 Adjusted R?= ,89100378 F(3,11)=39,148 p<,00000 Std.Error of estimate: 5,7357					
N=15	Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			26,10789	8,407584	3,10528	0,010009
x1	-0,126364	0,188376	-0,25180	0,375368	-0,67081	0,516179
x2	-0,744814	0,679027	-2,72767	2,486744	-1,09688	0,296124
x3	1,797200	0,666449	11,85602	4,396525	2,69668	0,020780

У нас

$$t_0 = 3{,}105278; \quad t_1 = -0{,}67081;$$

$$t_2 = -1{,}09688; \quad t_3 = 2{,}696681$$

табличне значення $t_{табл}(0{,}025{,}11) = 2{,}593097$

Оскільки $|t_0| > t_{\alpha/2,k}; \quad |t_1| < t_{\alpha/2,k}; \quad |t_2| < t_{\alpha/2,k}; \quad |t_3| > t_{\alpha/2,k}$

відповідно оцінки $\hat{a}_0, \hat{a}_3$ **є значущими**

а оцінки $\hat{a}_1, \hat{a}_2$ **не є значущими**

4 **Обчислимо коефіцієнти еластичності за формулою**

$$\alpha_i = \frac{\partial \hat{y}}{\partial x_i} \cdot \frac{\bar{x}_i}{\bar{y}}.$$

коефіцієнт еластичності є показником впливу зміни питомої ваги x_i на y у припущенні, що вплив інших факторів відсутній: показує, що регресанд y зміниться на α%, якщо фактор x зміниться на 1%

$$y = 26{,}108 - 0{,}252x_1 - 2{,}728x_2 + 11{,}856x_3$$

$$\frac{\partial \hat{y}}{\partial x_1} = -0{,}252$$

$$\bar{x}_1 = 23{,}416$$

$$\bar{y} = 43{,}08$$

$$\alpha_1 = \frac{\partial \hat{y}}{\partial x_1} \cdot \frac{\bar{x}_1}{\bar{y}} = -0{,}252 \cdot \frac{23{,}416}{43{,}08} \approx -0{,}137$$

У нас

$\alpha_1 = -0,137;\ \alpha_2 = -0,699;\ \alpha_3 = 1,24$

У нашому випадку він показує, що *прибуток підприємства зменшиться на 0,14 %*, якщо *інвестиції зростуть на 1%*, *прибуток підприємства зменшиться на 0,7 %*, якщо *витрати на рекламу зростуть на 1 %*, *прибуток підприємства збільшиться на 1,24 %*, якщо *заробітна плата зросте на 1 %*.

Загальна еластичність *Y* від усіх факторів x_i дорівнює:

$$\alpha = \sum_{i=1}^{m} \alpha_i$$

Цей показник свідчить, що на *α %* *зміниться y* , якщо одночасно збільшити на 1% всі фактори x_i)

У нас

α=0,394033.

5 **Обчислимо довірчі інтервали для математичного сподівання** $\hat{y}$ **і для кожного спостереження**

$$X_i^T = (1, x_{1(i)}, x_{2(i)}, x_{3(i)})$$

$$\left(\hat{y}_i - \Delta\hat{y}_i ; \hat{y}_i + \Delta\hat{y}_i\right),$$

де

$$\Delta\hat{y}_i = t_{\alpha/2,k} \cdot \hat{S}_u \cdot \sqrt{X_i^T (X^T X)^{-1} X_i}$$

де $\hat{S}_u$ — незміщена оцінка дисперсії залишків:

У нас

$$\hat{S}_u = 5,7357$$

Виконавши необхідні розрахунки, отримаємо довірчі зони регресії:

(23,430; 24,904)
(22,594; 22,604)
(24,730; 25,041)
....
(62,887; 63,558)
(68,229; 68,712)
(71,961;73,556)

6

Побудуємо довірчі інтервали для параметрів регресії.

$$\left(a_j - t_{\alpha/2,k}\sqrt{\sigma_u^2 c_{jj}}\ ;\ a_j + t_{\alpha/2,k}\sqrt{\sigma_u^2 c_{jj}}\right)$$

Довірчий інтервал при рівні надійності (1-*α*) є інтервал з випадково залежними межами і накриває істинне значення коефіцієнта регресії a_j *з рівнем довіри (1-α).*

c_{jj} діагональний елемент матриці $(X^T X)^{-1}$

У нас

N=15	Regression Summary for Dependent Variable: y (Данные) R= ,95622179 R?= ,91436011 Adjusted R?= ,89100378 F(3,11)=39,148 p<,00000 Std.Error of estimate: 5,7357					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			26,10789	8,407584	3,10528	0,010009
x1	-0,126364	0,188376	-0,25180	0,375368	-0,67081	0,516179
x2	-0,744814	0,679027	-2,72767	2,486744	-1,09688	0,296124
x3	1,797200	0,666449	11,85602	4,396525	2,69668	0,020780

Стандартні похибки

$$\sqrt{\sigma_u^2 c_{jj}}$$

$$t_{табл}(0{,}025;11) = 2{,}20$$

$$a_0 \in (7{,}85;\ 44{,}47) \qquad a_1 \in (-1{,}095;\ 0{,}539)$$

$$a_2 \in (-8{,}126;\ 2{,}801) \qquad a_3 \in (2{,}195;\ 21{,}44)$$

За допомогою пакету Excel.
Виконати команду **Сервис/Анализ данных**

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,956221789
R-квадрат	0,914360109
Нормированный R-кв	0,891003776
Стандартная ошибка	5,735708227
Наблюдения	15

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	3	3863,740536	1287,913512	39,14827207	3,6639E-06
Остаток	11	361,8818375	32,89834886		
Итого	14	4225,622373			

	Коэффициенты	*Стандарт ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	26,10788542	8,407584346	3,105277847	0,010009436	7,602907684	44,61286316	7,602907684	44,61286316
Переменная X 1	-0,251800474	0,375367538	-0,670810467	0,516178968	-1,077979273	0,574378325	-1,077979273	0,574378325
Переменная X 2	-2,727670903	2,486743611	-1,096884653	0,296123749	-8,200959455	2,745617649	-8,200959455	2,745617649
Переменная X 3	11,85602384	4,396524578	2,696680897	0,020779618	2,179333598	21,53271409	2,179333598	21,53271409

7

Обчислимо прогнозні значення і знайдемо межі довірчих інтервалів індивідуальних прогнозних значень і межі довірчих інтервалів для математичного сподівання (точковий та інтервальний прогнози).

а) для обчислення прогнозних значень $y_{pi} = Y_{np}$

у рівняння

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3$$

тобто

$$\hat{y} = 26{,}108 - 0{,}252 x_1 - 2{,}728 x_2 + 11{,}856 x_3$$

підставимо задані значення x_{pi}

У нас

Підставимо

$$x_{1np} = 48{,}82, \quad x_{2np} = 20{,}04, \quad x_{3np} = 10{,}25$$

одержимо

$$y_{np} = 80{,}68$$

б) знайдемо межі довірчих інтервалів індивідуальних прогнозованих значень за формулою:

$$\hat{Y}_{пр} - \Delta\hat{Y}_{пр} \le Y_{пр} \le \hat{Y}_{пр} + \Delta\hat{Y}_{пр},$$

де

$$\Delta\hat{Y}_{пр} = t_{\alpha/2}\hat{\sigma}_{u}\sqrt{1 + X_{пр}^{T}(X^{T}X)^{-1}X_{пр}}$$

У нас

$$\hat{Y}_{np} = 80,68 \qquad \sigma_u = 5,7357$$

$$X_{np} = \left(48,82;\ 20,04;\ 10,25\right)$$

тоді

$$Y_{np} \in (58,72;\ 102,64)$$

інтервальний прогноз **індивідуального значення**

в) знайдемо межі довірчих інтервалів для математичного сподівання значення y_{pi} за формулою:

$$\hat{Y}_{np} - \Delta_1 \le M(Y_{np}) \le \hat{Y}_{np} + \Delta_1,$$

де

$$\Delta_1 = t_{\alpha/2}\hat{\sigma}_u \sqrt{X_{np}^T (X^T X)^{-1} X_{np}}$$

У нас

$$\hat{Y}_{np} = 80{,}68 \qquad \sigma_u = 5{,}7357$$

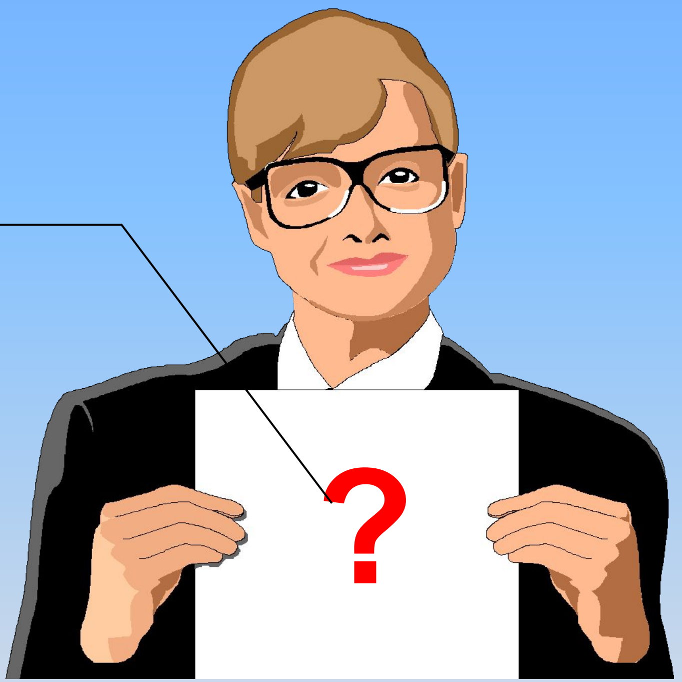
тоді

$$M(Y_{np}) \in (64{,}52;\ 96{,}83)$$

довірчий інтервал для *математичного сподівання*.

Завдання для самостійної роботи

Лугінін О.Є. Економетрія. Стор.123-132.
Приклад 6.3 розібрати розв ‘язок.



Експрес контроль

Знайти значення **критерію Фішера** для **парної** регресії, якщо $n=10+k$, $R^2=0,9t$,

де ***k***- номер по списку,

t- номер групи (1,2,3,4,5,6,7)