Лабораторна робота 6

Лінійна регресія

Створимо новий файл, в якому змінну VARI заповнимо послідовно значеннями від 0 до 10, змінну VAR2 – випадковими значеннями від 0 до 1, а змінну VAR3 задамо, як суму VAR1 + VAR2.

Виконаємо послідовність команд: *Statistics -> Basic Statistics/Tables -> Descriptive Statistics -> Prob.* &*Scatterplots* (див. рис. 1).

Descriptive Statistics: Spreadsheet1	?_X
Var1-Var3	Summary
Quick Advanced Normality Prob. & Scatterplots Categ. plots Options	Cancel
2D scatterplot with names S Normal probability plot	🔊 Options 👻
Use 3D scatterplot with names I Half-normal probability plot	
Categorized scatterplot	
Surface plot	
Scatterplot matrix	
	DF= © W4 O N4
	MD deletion
	Casewise Pairwise

Рис. 1

Як Variables виберемо VAR1-VAR3, натиснемо кнопку 2D Scatterplot.

У першому списку змінних вкажемо *VAR1*, в другому – *VAR3* і натиснемо кнопку *OK* (див. рис. 2).

На графіку, що з'явився, зображено пряму лінійної регресійної моделі для VAR3 через VAR1, а у верхній частині вікна бачимо рівняння лінійної регресії (див. рис. 3).

elect two var. lists (horizontal a	nd vertical vars in plots):	? ×
1-Var1 2-Var2 3-Var3	1-Var1 2-Var2 3-Var3	OK Cancel
Select All Spread Zoom First (horizontal) variable list: 1 1 Show appropriate variables only	Select All Spread Zoom Second (vertical) variable list: 3	

Рис. 2



Якщо у змінній *VAR3* замінити одне із значень, наприклад на 70, і побудувати графік знову, то побачимо, що рівняння регресії буде враховувати дане значення і з графіка буде очевидно, що 70 є викидом (див. рис. 4).

Натиснемо піктограму *Brushing*. У вікні, що з'явиться зробимо активними *Exclude* та *Box*, виділимо прямокутником значення (див. рис. 5) і натиснемо кнопку *Apply*. Виділене значення зникне з графіка і регресійна пряма змінить своє положення.

Для того, щоб значення викиду не виводилось у наступних графіках та не враховувалось при обчисленні регресійної формули, у змінній *VAR4* заповнимо всі значення одиницями, а те значення, що стоїть напроти викиду – нулем (див. рис. 6).



Рис. 4



1	1	2	3	4
	Var1	Var2	Var3	Var4
1	1	0,31204	1,31204	1
2	2	0,764626	2,764626	1
3	3	0,271319	3,271319	· · · · · · · · · · · · · · · · · · ·
4	4	0,107625	4,107625	
5	5	0,425613	70	(
6	6	0,681241	6,681241	3
7	7	0,95279	7,95279	
8	8	0,600465	8,600465	1
9	9	0,733896	9,733896	
10	10	0,239218	10,23922	

У вікні *Descriptive Statistics* натиснемо кнопку *Weight* (див. рис. 1). Оберемо змінну *VAR4*, перемкнемо *Status* на *On* та натиснемо *OK* (див. рис. 7). Тепер при аналізі викиди враховуватися не будуть.

Analysis/Graj	ph Case We	ights ? X			
Use Spreadsheet weights					
Weight variab	O Use weights for this Analysis/Graph only				
Status	Double-click	r on edit field to			
On choose from list of all variables. Values of selected					
	multipliers.	be used as case			
	OK	Cancel			

Рис. 7

Нехай маємо таку таблицю з даними:

Y	X	Ζ
40	100	10
50	200	20
50	300	10
70	400	30
65	500	20

65	600	20
80	700	30

де *Y* – врожайність, *X* – добрива, *Z* – опади. Потрібно знайти формулу багатофакторної лінійної регресії для *Y*:

$$Y = B_0 + B_1 X + B_2 Z$$
,

де *B_i* – невідомі коефіцієнти.

Виконаємо послідовність команд: *Statistics -> Multiple Regression*, як змінні оберемо Y – залежна, X і Z – незалежні (див. рис. 8). Натиснемо *OK*. Отримуємо результат, який зображено на рис. 9.

Multiple Linear Regression: Spreadsh	eet1 ?_>
Quick Advanced	с СК
🗩 Variables	Cancel
Dependent: Y Independent: X-Z	🔎 Options 🔻
	🗁 Open Data
	SELECT S
	Weighted moments
	OF=
	MD deletion
	Casewise
	C Pairwise C Mean substitution

Рис. 8

	ults: Spreadshee	t1			1	?_×
Multiple Regression)	Results					
Dependent: Y No. of cases: 7 Standard	Multiple adjusted error of estim	R = R?= R?= ate:	,99063942 ,98136646 ,97204969 2.314550249	F = 105, df = 2, p = ,00	.3333 .4)0347	
Intercept: 28,09523809	95 Std.Error:	2,491	482 t(4) = 11,277	p =	,0004
X beta=,55	94	Z bet	a=,491			
Inha for highlighting effects: 05	-					
1	Land I					OK
Quick Advanced Residuals/	assumptions/predictic	n				OK Cancel
Quick Advanced Residuals/	assumptions/predictic	n]				OK Cancel Options V
Quick Advanced Residuals/	assumptions/predictic	n]				OK Cancel Options V
Quick Advanced Residuals/	assumptions/predictio	n]				OK Cancel Options V

Рис. 9

В закладці *Quick* натиснемо кнопку *Summary: Regression results*. У вікні, що з'явилось (див. рис. 10), бачимо оцінки параметрів та допоміжну статистику. Обидві змінні значимі (виділені червоним). У третьому стовпці таблиці вказані оцінки для коефіцієнтів *B_i*. Отже,

Y=28.095+0.038**X*+0.833**Z*.

Якщо наші змінні попередньо стандартизувати, то у результаті такого регресійного

аналізу отримали б оцінки коефіцієнтів, які записані в першому стовпчику таблиці (зрозуміло, що $B_0=0$). Коефіцієнти з першого стовпця показують внесок у регресійну модель змінних X та Z.

Regression Summary for Dependent Variable: Y (Spreadsheet1) R= ,99063942 R?= ,98136646 Adjusted R?= ,97204969 F(2,4)=105,33 p<,00035 Std.Error of estimate: 2,3146						
	Beta	Std.Err.	В	Std.Err.	t(4)	p-level
N=7		of Beta		of B		
Intercept			28,09524	2,491482	11,27652	0,000352
Х	0,594430	0,091003	0,03810	0,005832	6,53197	0,002838
Z	0,491473	0,091003	0,83333	0,154303	5,40062	0,005690
Рис. 10						

Для того, щоб обчислити передбачуване значення для Y для заданих X та Z і побудувати 95% проміжок надійності, перейдемо у закладку *Residuals/assumptions/prediction* і натиснемо *Predict dependent variable* (див. рис. 11). У відповідні віконця вводимо значення змінних X та Z (див. рис. 12) і натискаємо OK.

Якщо незалежні змінні набувають одного і того ж значення його можна ввести у віконці *Common Value* і натиснути *Apply*.

У вікні результатів аналізу (див. рис. 13) бачимо передбачуване значення *Y*, верхню і нижню межу надійного проміжку для цього значення.

	Spreadsheet1	?_)
Multiple Regression Resul	ts	
Dependent: Y No. of cases: 7 Standard errc Intercent: 28.095238095 8	Multiple R = ,99063942 F = 1(R?= ,98136646 df = adjusted R?= ,97204969 p = or of estimate: 2,314550249 5td Error: 2,491482 t(4) = 11.2'	05,3333 2,4 ,000347 77 p = .0004
X beta=,594	Z beta=,491	
(significant betas are hig oha for highlighting effects: 0.05 💌	phlighted)	<u>∎</u> ■ OK
Quick Advanced Residuals/assum	ptions/prediction	Cancel
		Caricor
Perform residual analysis Descriptive statistics Code generator		Diptions •
Perform residual analysis Descriptive statistics Code generator	Predict values ? Predict dependent variable @ Compute confidence limits Alpha: Compute prediction limits .05	Difference of the second secon
 Perform residual analysis Descriptive statistics Code generator 	Predict values Predict dependent variable Compute confidence limits Alpha: Compute prediction limits .05 Э Рис. 11	



Рис. 12

	Predicting Values for (Spreadsheet1) variable: Y				
	B-Weight	Value	B-Weight		
√ariable			* Value		
X	0,038095	250,0000	9,52381		
Z	0,833333	30,0000	25,00000		
Intercept			28,09524		
Predicted			62,61905		
-95,0%CL			55,99196		
+95,0%CL			69,24614		
D 10					

Рис. 13

Якщо ж потрібно подивитись як розподілені залишки, то натиснемо кнопку *Perform residual analysis*. У вікні, що з'явилось (див. рис. 14), зібрані різні методи для аналізу залишків регресійної моделі.

🖉 Residual Analysis: Sp	readsheet1			<u>? _</u> X
Dependent: Y No. of cases: 7 Standard en Intercept: 28,09523	Multiple R : R?: adjusted R?: cror of estimate: 8095 Std.Error:	,99063942 ,98136646 ,97204969 2,314550249 2,491482 t(F = 105,3333 df = 2,4 p = ,000347 4) = 11,277	p ≺ ,0004 ৳ ±
Quick Advanced Residu	als Predicted Scatter	plots Probability pl	lots Outliers Save	Cancel
Normal plot of re	siduals			Doptions -



Наприклад, натиснувши Normal plot of residuals отримаємо Q-Q графік, на якому видно наскільки залишки узгоджуються з нормальним законом розподілу (див. рис. 15).



Контрольні запитання

1. Як створити новий документ з трьома змінними? Як заповнити змінну послідовно значеннями 1, 2, 3, 4, 5 і т.д. Як заповнити змінну випадковими величинами від 0 до 1? (Опишіть послідовність дій)

2. Як виключити з розрахунку і побудови графіків викиди?

3. Як можна забезпечити, щоб значення викиду не виводилось у наступних графіках та не враховувалось при обчисленні регресійної формули?

4. Як знайти формулу багатофакторної лінійної регресії? У чому полягає відмінність між однофакторною і багтофакторною регресією?