

Лабораторна робота №7

Багатофакторна регресія

Відкриємо файл *Job_prof.sta* (див. рис. 1). У файлі вказано бали, отримані претендентами під час тестування при прийнятті на посаду (у перших чотирьох стовпцях), та оцінка професійної здатності претендентів (у п'ятому стовпці) після закінчення випробувального терміна. Нам потрібно знайти лінійну багатофакторну регресійну модель залежності оцінки професійної здатності від оцінок за тести. Завантажимо модуль *Multiple Regression: Statistics -> Multiple Regression*.

Job proficiency data set from Neter, Was:

	1	2	3	4	5
	TEST1	TEST2	TEST3	TEST4	JOB_PROF
1	88	110	100	87	88
2	62	97	99	100	80
3	110	107	103	103	96
4	101	117	93	95	76
5	100	101	95	88	80
6	78	85	95	84	73
7	120	77	80	74	58
8	105	122	116	102	116
9	112	119	106	105	104
10	120	89	105	97	99
11	87	81	90	88	64
12	133	120	113	108	126
13	140	121	96	89	94
14	84	113	98	78	71
15	106	102	109	109	111
16	109	129	102	108	109
17	104	83	100	102	100
18	150	118	107	110	127
19	98	125	108	95	99
20	120	94	95	90	82
21	74	121	91	85	67
22	96	114	114	103	109
23	104	73	93	80	78
24	94	121	115	104	115
25	91	129	97	83	83

Рис. 1

Натиснувши *Variables*, обираємо *Job-Prof*, як залежну змінну, а як незалежні змінні вибираємо перші чотири змінні (див. рис. 2). Двічі натискаємо *OK*. Результати регресійного аналізу зображені на рисунку 3. Всі змінні, окрім другої, є значимими (виділені червоним).

У закладці *Quick* натиснемо *Summary: Regression result* (див. рис. 3). У вікні, що з'явилося (див. рис. 4), бачимо результати аналізу: у третьому стовпці – коефіцієнти багатофакторної лінійної регресійної моделі, а в першому стовпці – коефіцієнти цієї ж регресійної моделі для стандартизованих змінних.

Проаналізувавши результати прийдемо до висновку, що *Test 2* досить мало впливає на оцінку професійної здатності претендентів: відповідний коефіцієнт у першому стовбці становить 0,043. Тому можливо є доречним взагалі вилучити *Test 2* з регресійної моделі.

Select dependent and independent variable lists: ? X

1-TEST1
2-TEST2
3-TEST3
4-TEST4
5-JOB_PROF

1-TEST1
2-TEST2
3-TEST3
4-TEST4
5-JOB_PROF

OK
Cancel

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Select All Spread Zoom

Select All Spread Zoom

Dependent var. (or list for batch):
5

Independent variable list:
1-4

☐ Show appropriate variables only

Рис. 2

Multiple Regression Results: Job_prof ? - X

Multiple Regression Results

Dependent: JOB_PROF Multiple R = ,98127051 F = 129,7411
R² = ,96289181 df = 4,20
No. of cases: 25 adjusted R² = ,95547018 p = ,000000
Standard error of estimate: 4,098643206
Intercept: -124,3818206 Std. Error: 9,941063 t(20) = -12,51 p = ,0000

TEST1 beta=,309 TEST2 beta=,043 TEST3 beta=,595
TEST4 beta=,286

(significant betas are highlighted)

Alpha for highlighting effects: .05 OK

Quick Advanced Residuals/assumptions/prediction

Summary Summary: Regression results Options

Рис. 3

Regression Summary for Dependent Variable: JOB_PROF (Jo

R= ,98127051 R²= ,96289181 Adjusted R²= ,95547018
F(4,20)=129,74 p<,00000 Std. Error of estimate: 4,0986

	Beta	Std. Err. of Beta	B	Std. Err. of B	t(20)	p-level
N=25						
Intercept			-124,382	9,941063	-12,5119	0,000000
TEST1	0,309042	0,045951	0,296	0,043971	6,7254	0,000002
TEST2	0,042992	0,050408	0,048	0,056617	0,8529	0,403826
TEST3	0,595438	0,074813	1,306	0,164091	7,9591	0,000000
TEST4	0,285723	0,072524	0,520	0,131943	3,9397	0,000810

Рис. 4

Проведемо спочатку аналіз залишків побудованої моделі. У закладці *Residuals/assumptions/prediction*, натиснемо кнопку *Perform residual analysis*. У вікні, що з'явилося, обираємо *Residuals* -> *Casewise plot of residuals* (див. рис. 5).

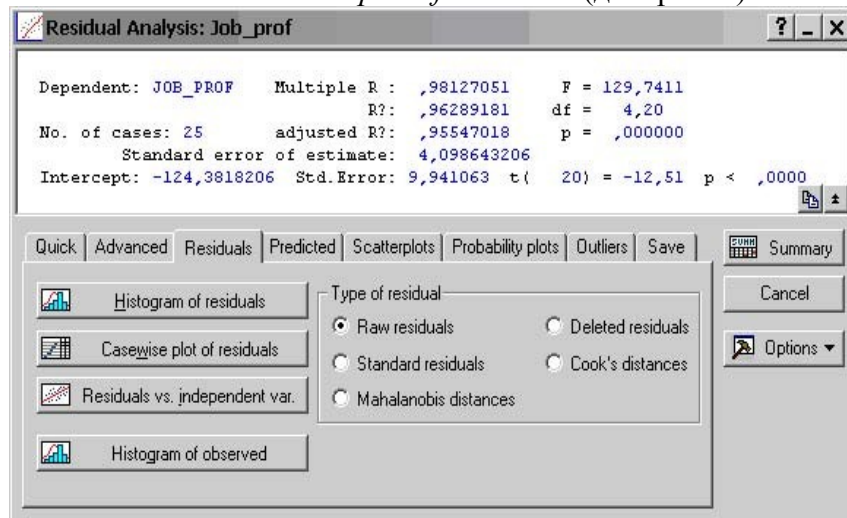


Рис. 5

Так ми перевіримо, чи не виходять залишки за межі 3δ . Отримаємо таблицю (див. рис. 6), в якій знаком «*» вказано де знаходиться залишок у інтервалі $(-3\delta; 3\delta)$. У правій частині таблиці міститься додаткова інформація про залишки.

Бачимо, що залишки лежать у проміжку $(-2\delta; 2\delta)$ і їхнє середнє і медіана дорівнюють нулю.

Потім оберемо закладку *Outliers* (див. рис. 7). Натиснемо *Casewise plot outliers* та упевнимися, що викидів немає – з'явиться відповідне повідомлення (див. рис. 8).

Raw Residual (Job_prof)									
Dependent variable: JOB_PROF									
Case	Raw Residuals					Observed Value	Predicted Value	Residual	Standard Residual
	-3σ	-2σ	-σ	0	+σ				
1	.	.	.	.	.	88,0000	82,1078	5,8922	0,52533
2	.	.	.	*	.	80,0000	78,8141	1,1859	0,02087
3	.	.	*	.	.	96,0000	101,3753	-5,3753	-0,48141
4	.	.	*	.	.	76,0000	81,9779	-5,9779	-0,53833
5	.	.	.	*	.	80,0000	78,8828	1,1172	0,02857
6	.	.	.	*	.	73,0000	70,5251	2,4749	0,20384
7	.	.	.	*	.	50,0000	57,7709	-7,7709	-0,64666
8	.	.	*	.	.	116,0000	117,0783	-1,0783	-0,08537
9	.	.	*	.	.	104,0000	107,5938	-3,5938	-0,35039
10	.	.	*	.	.	98,0000	102,9584	-4,9584	-0,49584
11	.	.	*	.	.	64,0000	68,5427	-4,5427	-0,45427
12	.	.	*	.	.	128,0000	124,4639	3,5361	0,35361
13	.	.	*	.	.	94,0000	94,5035	-0,5035	-0,05035
14	.	.	*	.	.	71,0000	74,4506	-3,4506	-0,34506
15	.	.	*	.	.	111,0000	110,9059	0,0941	0,00941
16	.	.	.	*	.	109,0000	103,4349	5,5651	0,55651
17	.	.	.	*	.	100,0000	94,0041	5,9959	0,59959
18	.	.	.	*	.	127,0000	122,5982	4,4018	0,44018
19	.	.	*	.	.	99,0000	101,0672	-2,0672	-0,20672
20	.	.	*	.	.	82,0000	86,4880	-4,4880	-0,44880
21	.	.	*	.	.	67,0000	66,3763	0,6237	0,06237
22	.	.	*	.	.	109,0000	111,9392	-2,9392	-0,29392
23	.	.	*	.	.	78,0000	72,9432	5,0568	0,50568
24	.	.	*	.	.	83,0000	78,5984	4,4016	0,44016
25	.	.	*	.	.	68,0000	57,7709	10,2291	1,02291
Minimum	.	.	*	.	.	127,0000	124,4639	2,5361	0,25361
Maximum	.	.	*	.	.	92,0000	92,0000	0,0000	0,00000
Mean	.	.	*	.	.	84,0000	84,0000	0,0000	0,00000
Median	.	.	*	.	.	84,0000	84,0000	0,0000	0,00000

Рис. 6

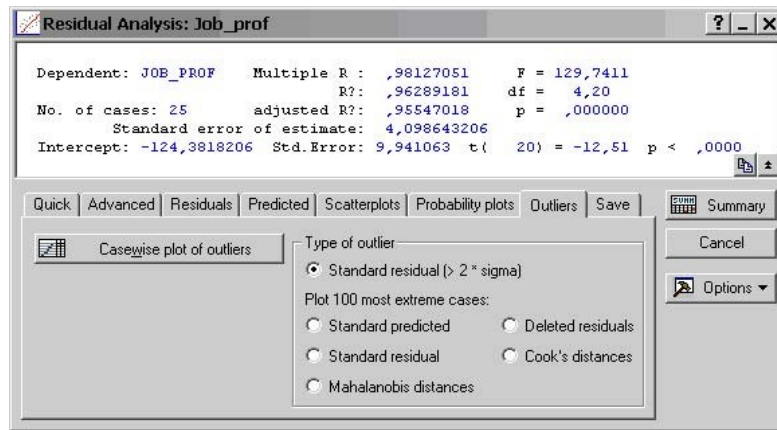


Рис. 7

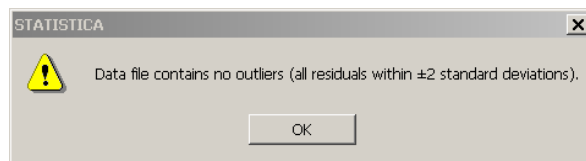


Рис. 8

На основі цих результатів можна вважати, що багатфакторна регресійна модель достатньо добре описує наші дані.

Цікаво дослідити, що відбудеться, якщо вилучити змінну *Test 2*, яка досить мало впливає на оцінку професійної здатності претендентів, з регресійної моделі. Розглянемо, як автоматизовано процес знаходження змінних, які дають малий внесок у регресійну модель, у пакеті *Statistica*.

У вікні *Multiple Regression* у закладці *Advanced* відмітимо *Advanced options (stepwise or ridge regression)* та натиснемо *OK* (див. рис. 9).

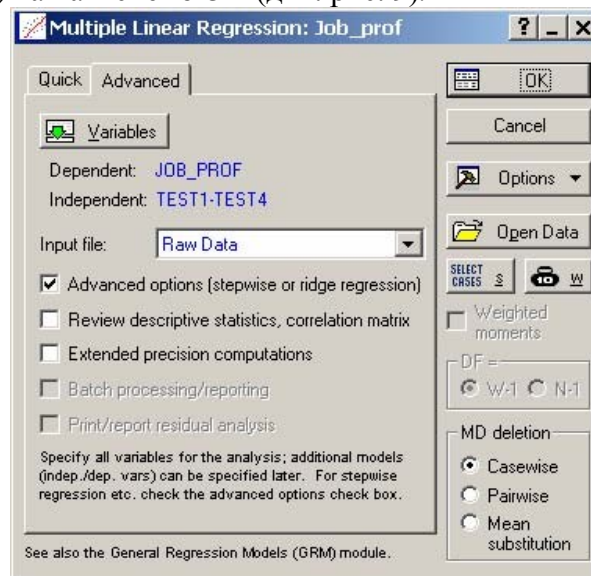


Рис. 9

У закладці *Stepwise* вкажемо метод *Forward stepwise* – змінні будуть введені у регресійну модель по одній. У *Display results* вкажемо покроковий вивід результатів – *At each step* (див. рис. 10). Тобто ми будемо здійснювати покрокову регресію, з виводом результатів після кожного кроку. У полі *F to enter* вказуємо значення 1, а у полі *F to remove* вказуємо 0.01. Ці два числа визначають верхню та нижню межі проміжку для значимості внеску у регресійну модель змінних. Якщо значимість змінної потрапляє в цей проміжок, то включаємо її до регресійної моделі, інакше – відкидаємо.

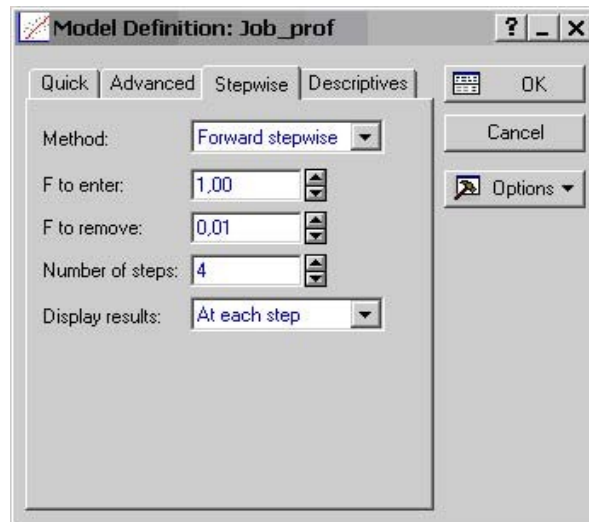


Рис. 10

Оскільки маємо чотири незалежні змінні, то кількість кроків множинної регресії *Number of steps* досить вказати рівною чотирьом (після кожного кроку до моделі може включатись не більше однієї змінної). Натискаємо кнопку *OK*.

У вікні, що з'явилося, бачимо, що в моделі ще немає жодної змінної (див. рис. 11). Натискаємо *Next*. З'явився перший коефіцієнт та перша, вибрана до рівняння регресії змінна (див. рис. 12). Натискаємо кнопку *Next*, доки на цій кнопці не з'явиться напис *OK*. Це означатиме, що процедуру вибору змінних до регресійної моделі закінчено і всі змінні значимість внеску яких знаходиться у вказаних межах увійшли до рівняння регресії. Бачимо, що змінна *Test 2* не увійшла у нову багатофакторну регресійну модель (див. рис. 13).

Далі можемо подивитися результати. Для цього натиснемо на закладці *Advanced*, кнопку *Stepwise Regression Summary* (див. рис. 14). У вікні, що з'явилося (див. рис. 15), бачимо, статистику внеску обраних змінних і порядок включення їх у регресійну модель. Якщо натиснемо на закладці *Advanced*, кнопку *Summary: Regression results* (див. рис. 14), то у вікні, що з'явилося (див. рис. 16), побачимо з якими коефіцієнтами змінні увійшли до регресійної моделі.

Використовуючи отриману інформацію (рис.15 і 16) можемо порівняти стару і нову регресійну моделі. Бачимо, що вони майже не відрізняються. Отже, внесок змінної *Test 2* дійсно був незначним.

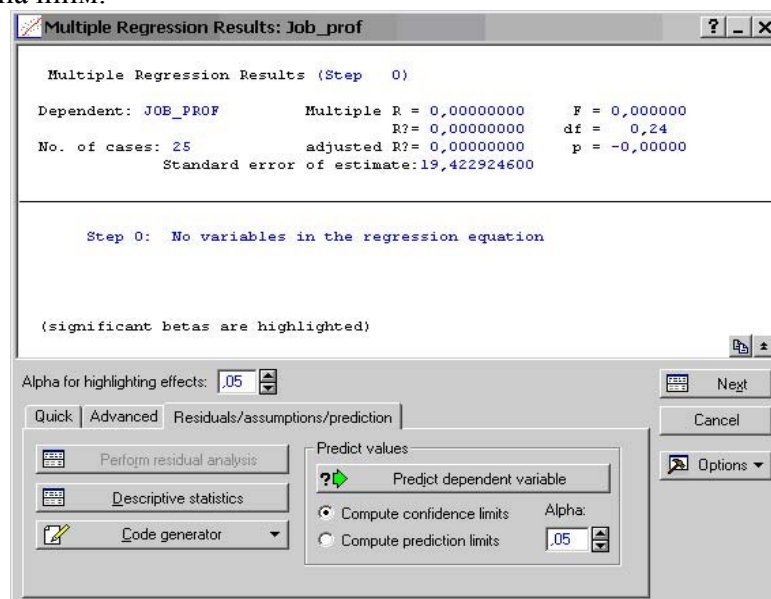


Рис. 11

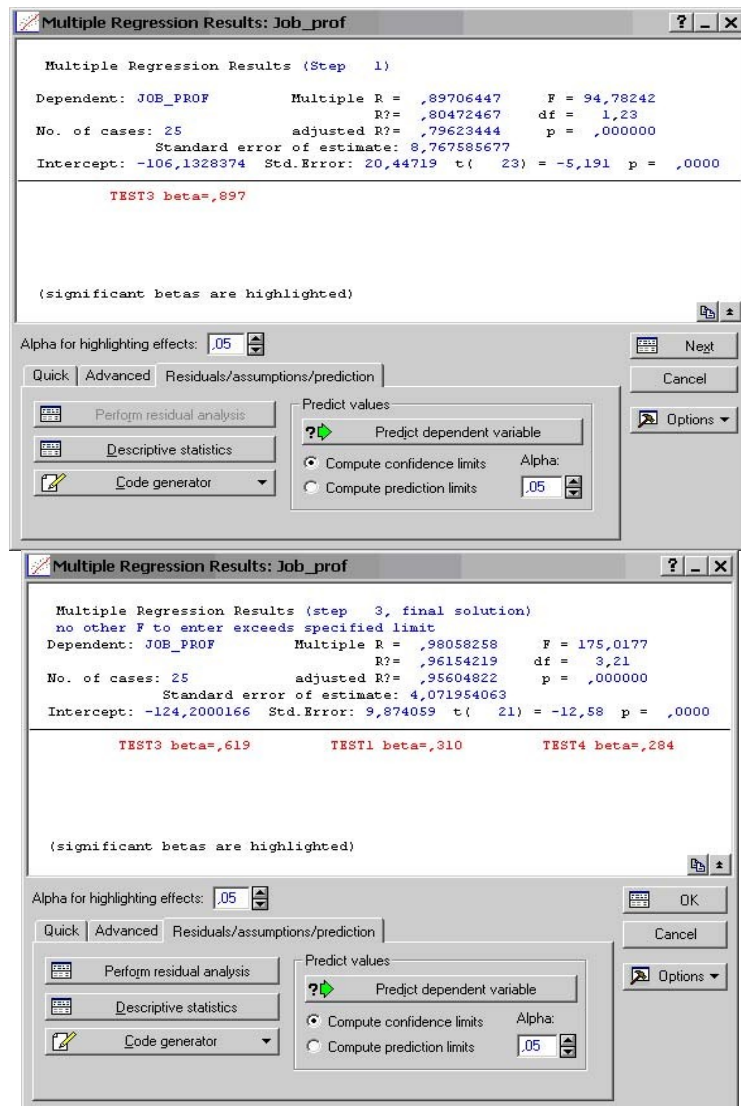
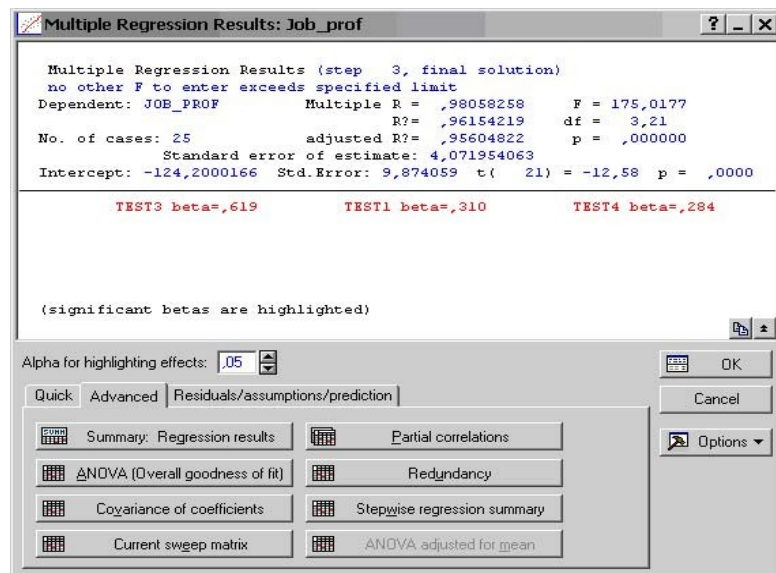


Рис. 13



Summary of Stepwise Regression; DV: JOB_PROF (Job_prof)

Variable	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to entr/rem	p-level	Variables included
TEST3	1	0,897064	0,804725	0,804725	94,78241	0,000000	1
TEST1	2	0,965917	0,932996	0,128271	42,11609	0,000002	2
TEST4	3	0,980583	0,961542	0,028547	15,58793	0,000735	3

Рис. 15

Regression Summary for Dependent Variable: JOB_PROF (Job_prof)

R = .98058258 R² = .96154219 Adjusted R² = .95604822
F(3,21) = 175.02 p < .00000 Std. Error of estimate: 4.0720

N=25	Beta	Std. Err. of Beta	B	Std. Err. of B	t(21)	p-level
Intercept			-124,200	9,874059	-12,5784	0,000000
TEST3	0,618670	0,069224	1,357	0,151832	8,9373	0,000000
TEST1	0,309670	0,045646	0,296	0,043679	6,7841	0,000001
TEST4	0,284405	0,072035	0,517	0,131054	3,9482	0,000735

Рис. 16

Якщо є мультиколінеарність змінних (наприклад, в кількох стовпцях містяться дані про ціну одного і того ж товару в різних валютах), то очевидно, що для регресійної моделі слід взяти не всі із таких змінних.

Якщо для обробки таблиці мультиколінеарних даних так як і раніше скористатися стандартним алгоритмом, то з'явиться повідомлення про помилку (див. рис. 17).

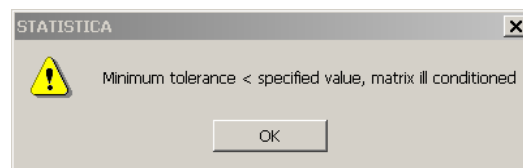


Рис. 17

У цьому разі для регресійного аналізу потрібно вибирати метод *Forward stepwise* (який дозволяє вводити у модель по одній змінній), а не *Standard*.

Контрольні запитання

1. На рис. 4. представлені результати аналізу. Яким чином з рисунка можна визначити які змінні несуттєво впливають на регресійну модель?
2. На основі яких результатів можна вважати, що багатофакторна регресійна модель достатньо добре описує дані?
3. Скільки незалежних змінних розглядається в прикладі *Job_prof.sta* ?
4. Наведіть приклад мультиколінеарності змінних.